Exploring the value of Big Data analysis of Twitter tweets and share prices

A Thesis submitted by

Peter Wlodarczak, B.Sc. Computer Science, MBA

For the award of

Doctor of Philosophy,

2017

Abstract

Over the past decade, the use of social media (SM) such as Facebook, Twitter, Pinterest and Tumblr has dramatically increased. Using SM, millions of users are creating large amounts of data every day. According to some estimates ninety per cent of the content on the Internet is now user generated. Social Media (SM) can be seen as a distributed content creation and sharing platform based on Web 2.0 technologies. SM sites make it very easy for its users to publish text, pictures, links, messages or videos without the need to be able to program. Users post reviews on products and services they bought, write about their interests and intentions or give their opinions and views on political subjects. SM has also been a key factor in mass movements such as the Arab Spring and the Occupy Wall Street protests and is used for human aid and disaster relief (HADR).

There is a growing interest in SM analysis from organisations for detecting new trends, getting user opinions on their products and services or finding out about their online reputation. Companies such as Amazon or eBay use SM data for their recommendation engines and to generate more business. TV stations buy data about opinions on their TV programs from Facebook to find out what the popularity of a certain TV show is. Companies such as Topsy, Gnip, DataSift and Zoomph have built their entire business models around SM analysis.

The purpose of this thesis is to explore the economic value of Twitter tweets. The economic value is determined by trying to predict the share price of a company. If the share price of a company can be predicted using SM data, it should be possible to deduce a monetary value. There is limited research on determining the economic value of SM data for “nowcasting”, predicting the present, and for forecasting. This study aims to determine the monetary value of Twitter by correlating the daily frequencies of positive and negative Tweets about the Apple company and some of its most popular products with the development of the Apple Inc. share price. If the number of positive tweets about Apple increases and the share price follows this development, the tweets have predictive information about the share price.

A literature review has found that there is a growing interest in analysing SM data from different industries. A lot of research is conducted studying SM from various perspectives. Many studies try to determine the impact of online marketing campaigns or try to quantify the value of social capital. Others, in the area of behavioural economics, focus on the influence of SM on decision-making. There are studies trying to predict financial indicators such as the Dow Jones Industrial Average (DJIA). However, the literature review has indicated that there is no study correlating sentiment polarity on products and companies in tweets with the share price of the company.

The theoretical framework used in this study is based on Computational Social Science (CSS) and Big Data. Supporting theories of CSS are Social Media Mining (SMM) and sentiment analysis. Supporting theories of Big Data are Data Mining (DM) and Predictive

Analysis (PA). Machine learning (ML) techniques have been adopted to analyse and classify the tweets.

In the first stage of the study, a body of tweets was collected and pre-processed, and then analysed for their sentiment polarity towards Apple Inc., the iPad and the iPhone. Several datasets were created using different pre-processing and analysis methods. The tweet frequencies were then represented as time series. The time series were analysed against the share price time series using the Granger causality test to determine if one time series has predictive information about the share price time series over the same period of time. For this study, several Predictive Analytics (PA) techniques on tweets were evaluated to predict the Apple share price.

To collect and analyse the data, a framework has been developed based on the LingPipe (LingPipe 2015) Natural Language Processing (NLP) tool kit for sentiment analysis, and using R, the functional language and environment for statistical computing, for correlation analysis. Twitter provides an API (Application Programming Interface) to access and collect its data programmatically.

Whereas no clear correlation could be determined, at least one dataset was showed to have some predictive information on the development of the Apple share price. The other datasets did not show to have any predictive capabilities. There are many data analysis and PA techniques. The techniques applied in this study did not indicate a direct correlation. However, some results suggest that this is due to noise or asymmetric distributions in the datasets.

The study contributes to the literature by providing a quantitative analysis of SM data, for example tweets about Apple and its most popular products, the iPad and iPhone. It shows how SM data can be used for PA. It contributes to the literature on Big Data and SMM by showing how SM data can be collected, analysed and classified and explore if the share price of a company can be determined based on sentiment time series. It may ultimately lead to better decision making, for instance for investments or share buyback.

Certification of Thesis

This thesis is entirely the work of Peter Wlodarczak except where otherwise acknowledged. The work is original and has not previously been submitted for any other award, except where acknowledged.
Student and supervisors signatures of endorsement are held at USQ.

Principal Supervisor

Dr. Mustafa Ally

Associate Supervisor

Prof Jeffrey Soar

Acknowledgements

I would like to thank my supervisors for their valuable input. Special thanks go to my principal supervisors, Dr. Mustafa Ally and Prof. Dr. Jeffrey Soar, for their encouragement and always constructive feedback. I am grateful for their patience, guidance and insight throughout this research project and for them proof-reading the thesis.
I am also thankful to my friends and family for their support despite the fact that I had too little time for them during this research project.

Table of content

1 Introduction 1
1.1 Background 1
1.2 Justification for the research 2
1.3 Research methodology 3
1.3.1 Data conditioning phase 5
1.3.2 Predictive analysis phase 5
1.4 Delimitation of the scope and key assumptions 7
1.5 Key definitions and terminologies 9
1.5.1 Computational social science (CSS) 9
1.5.1.1 Social media mining (SMM) 10
1.5.1.2 Sentiment analysis 10
1.5.2 Big Data 11
1.5.2.1 Data mining (DM) 12
1.5.2.2 Predictive analytics (PA) 12
1.5.3 Artificial intelligence (AI) 13
1.5.3.1 Machine learning (ML) 13
1.5.3.2 Data classification 14
1.5.4 Theoretical framework 15
1.6 Publication list 16
1.7 Structure of the thesis 17
2 Literature review 19
2.1 Introduction 19
2.1.1 Computational social science 21
2.1.1.1 Social media mining 21
2.1.1.2 Sentiment analysis (SA) 25
2.1.1.3 Quantifying social media data 26
2.1.1.4 Frameworks for computational analysis 28
2.1.2 Big Data 32
2.1.2.1 Data mining (DM) 34
2.1.2.2 Predictive analysis (PA) 37
2.2 A framework for analysing Twitter data 40
2.3 Research question and issues 41
2.4 Summary 42
3 Research methodology 44
3.1 Introduction 44
3.1.1 Data conditioning phase 44
3.1.1.1 Data collection of Twitter data 44
3.1.1.1.1 Datasets 49
3.1.1.2 Data pre-processing 50
3.1.2 Predictive analysis phase 52
3.1.2.1 Data classification 52
3.1.2.2 Correlations 57
3.2 Data analysis framework 59
3.3 Justification for the paradigm and methodology 60
3.4 Ethical considerations 62

3.5 Conclusions ... 64
4 Data analysis ... 65
4.1 Introduction ... 65
4.2 Data mining ... 65
4.2.1 Data conditioning phase ... 65
4.2.1.1 Access method ... 67
4.2.1.1.1 Queries ... 69
4.2.1.2 Data collection ... 70
4.2.1.3 Data pre-processing ... 74
4.2.1.4 Natural language processing ... 75
4.2.1.4.1 Data deduplication ... 77
4.2.1.4.2 Basic subjectivity analysis ... 78
4.2.1.5 Classifier evaluation ... 79
4.2.2 Predictive analysis phase ... 85
4.2.2.1 Multigram language model ... 88
4.2.2.2 Naïve Bayes ... 91
4.2.2.3 Logistic regression ... 96
4.2.2.4 Summary ... 99
4.2.2.5 Granger causality test ... 99
4.3 Results ... 103
4.3.1 Apple datasets ... 103
4.3.2 iPad datasets ... 104
4.3.3 iPhone datasets ... 105
4.4 Conclusions ... 106
5 Conclusions and implications ... 112
5.1 Introduction ... 112
5.2 Conclusions about the research problem ... 112
5.3 Implications for theory ... 114
5.4 Implications for policy and practise ... 117
5.5 Limitations ... 119
5.6 Further research ... 121
5.7 Conclusions ... 124
6 Appendices ... 127
6.1 References ... 127
6.2 Queries ... 147
6.3 Results ... 148
6.3.1 Results of the Granger causality test ... 148
6.3.1.1 Apple unclassified ... 148
6.3.1.2 Apple classified naïve Bayes ... 149
6.3.1.3 Apple classified Logistic Regression ... 150
6.3.1.4 iPad unclassified ... 151
6.3.1.5 iPad classified naïve Bayes ... 152
6.3.1.6 iPad classified Logistic Regression ... 153
6.3.1.7 iPhone unclassified ... 154
6.3.1.8 iPhone classified naïve Bayes ... 154
6.3.1.9 iPhone classified Logistic Regression ... 156
6.4 Abbreviations ... 157

List of tables

Table 1-1: Publication list 17
Table 2-1: Literature search terms 20
Table 2-2: Computational social science literature overview 31
Table 2-3: Big Data literature overview 40
Table 4-1: Twitter query for Apple 69
Table 4-2: Excel export of AAPL quotes 73
Table 4-3: Data deduplication percent per term 78
Table 4-4: Summary statistics for basic subjectivity analysis 79
Table 4-5: Confusion matrices 80
Table 4-6: Percentage of objective statements 84
Table 4-7: Test sentences 89
Table 4-8: Test sentence results with 737 training tweets 90
Table 4-9: Summary statistics for Language Model classifier 90
Table 4-10: Confusion matrices 91
Table 4-11: Test sentence results with 737 training tweets 93
Table 4-12: Naïve Bayes classifier 94
Table 4-13: Confusion matrices 94
Table 4-14: Naïve Bayes classifier using 10-fold cross-validation 95
Table 4-15: Confusion matrices 95
Table 4-16: Logistic regression without cross-validation 97
Table 4-17: Confusion matrices 97
Table 4-18: Logistic regression with 10-fold cross-validation 98
Table 4-19: Confusion matrices 98
Table 4-20: Logistic regression using character n-grams 98
Table 4-21: Confusion matrices 98
Table 4-22: Apple unclassified 103
Table 4-23: Apple classified 104
Table 4-24: iPad unclassified 104
Table 4-25: iPad classified 105
Table 4-26: iPhone unclassified 105
Table 4-27: iPhone classified 106
Table 6-1: Apple unclassified 149
Table 6-2: Apple classified naïve Bayes 150
Table 6-3: Apple classified Logistic Regression 151
Table 6-4: iPad unclassified 151
Table 6-5: iPad classified naïve Bayes 153
Table 6-6: iPad classified Logistic Regression 154
Table 6-7: iPhone unclassified 154
Table 6-8: iPhone classified naïve Bayes 155
Table 6-9: iPhone classified Logistic Regression 156

List of figures

Figure 1-1: Analysis phases 4
Figure 1-2: Data conditioning and pretictive analysis stages 7
Figure 1-3: Theoretical framework 15
Figure 3-1: Twitter data query output 47
Figure 3-2: Twitter search 48
Figure 3-3: Relevance filtering of tweet 50
Figure 3-4: Machine learning cycle 55
Figure 3-5: Binary sentiment classification time series 57
Figure 3-6: Tweets about Apple Inc. and AAPL share price 58
Figure 3-7: Data analysis framework 60
Figure 4-1: Twitter data mining steps 65
Figure 4-2: Number of tweets about Apple, the iPhone and the iPad 66
Figure 4-3: Tweets imported in NVivo 71
Figure 4-4: AAPL quotes over a period of 2 month 73
Figure 4-5: Precision Recall 82
Figure 4-6: Receiver Operating Characteristic 84
Figure 4-8: Naive Bayes classifier 94
Figure 4-9: Sample grangertest output 101
Figure 4-10: R time series plot of Apple tweet frequency and share price 102
Figure 4-11: Frequencies of tweets 109
Figure 4-12: Apple Q4 2015 results (Apple Inc. 2016) 110

List of equations

Equation 3-1: Classification probability 55
Equation 4-1: Jaccard index 77
Equation 4-3: F1 score 81
Equation 4-4: Precision 82
Equation 4-5: Recall 82
Equation 4-6: Kappa coefficient 83
Equation 4-7: Observer agreement 83
Equation 4-8: Category probability 92
Equation 4-9: Granger causality test 99

1 Introduction

1.1 Background

Opinions are central to almost all human activities because they are key influences of our behaviours (Liu 2012, p. 2). People want to know other people's opinions whenever they need to make a decision about choosing a holiday destination, making a voting decision in an election, selecting a restaurant or buying a new car. Emotions towards products or services influence our purchasing decisions. Behavioural economics tells us that emotions can profoundly affect individual behaviour and decision-making (Bollen, Mao & Zeng 2010, p. 1).

In the past, people turned to friends and family whenever they wanted to know an individual's opinion. When an organisation wanted to obtain public opinions, it had to conduct surveys, focus groups or polls. With the unprecedented rise of the internet and especially Web 2.0 technologies, sharing opinions and views has become increasingly easy. Web 2.0 is a technology shifting the Web to turn it into a more participatory platform, in which people not only consume content (via downloading) but also contribute and produce new content (via uploading) (Darwish & Lakhtaria 2011, p. 204).

SM has exploded as a category of online discourse where people create content, share it, bookmark it and network at a prodigious rate (Asur & Huberman 2010, p. 492). SM has been increasingly used in a wide range of domains, such as political campaigns (for example, presidential elections), mass movements (for example, organizing Occupy Wall Street movements, Arab Spring), as well as disaster and crisis response and relief coordination (Gundecha & Liu 2012, p. 12).

With the explosive growth in the availability and use of Social Media (SM) (e.g., reviews, forum discussions, blogs, microblogs, comments, and postings in social network sites) on the Web, individuals and organisations are increasingly using the content in these media for decision making (Liu 2012, p. 2). Customer reviews make a significant impact on the purchasing decision of potential customers (Smith et al. 2011, p. 20). As such they can influence the propensity of customers to buy products or services or influence the choices for holiday destinations or restaurants. If reviews written in the past influence future actions of customers, SM data has the potential to be used for forecasting, predicting the future, or nowcasting, that is, to predict the present in real-time, for instance the rise or fall of share prices.

Organisations have realised the potential of SMM (Social Media Mining) in extracting actionable patterns that can be beneficial for business, users and consumers (Gundecha & Liu 2012, p. 1). SM is analysed for its potential to predict sales volumes, to identify customers who are likely to switch the supplier or to determine customer satisfaction. If opinions can be used to make predictions about sales volumes, they might also have predictive information about other financial indicators.

Not surprisingly, the need for studying and understanding the social phenomena underlying such communities has recently given rise to a new field of work, known as Computational Social Science, which is materializing at the crossroads of computer science and the social sciences (Smith et al. 2011, p. 1).

1.2 Justification for the research

The increase in the use of SM has led many social scientists to examine whether specific patterns in the streams of tweets might be able to predict real-world outcomes (Bollen,

Mao & Zeng 2010; Barberá & Rivero 2013; Gerber 2014; Burnap et al. 2015; Tsakalidis et al. 2015). Predictive Analysis (PA) of SM data is a recent area of research (Hong Keel, Dennis & Yuan 2014; Burnap et al. 2015; Tsakalidis et al. 2015). Despite attracting an increasing number of researchers; many areas are still unexplored. The potential of analysing SM has been recognized but the literature review in chapter 2 Literature review has identified gaps in exploring the influence of SM on share price developments. Also there are many different techniques for SMM and PA. Some have proven to be very effective for certain tasks but have performed less well in other environments. Ipso facto, assessing the best performing methods for a specific problem is an important SMM task since often it is not possible to tell a priori which method will yield the best results (Witten, Frank & Hall 2011). Analysing and evaluating different SMM and PA techniques and algorithms can thus lead to better outcomes, and research in this area has the potential to improve the analysis of SM and the accuracy of PA (Huang et al. 2013).

1.3 Research methodology

Data analysis, in the context of this study, comprises two phases. The first phase is the data conditioning phase where data is collected, and passed through pre-processing steps such as relevance filtering or data deduplication. The main purpose of the data conditioning phase is the transformation of noisy raw SM data into high-quality data that will enable the computation of predictor variables (Kalampokis, Tambouris & Tarabanis 2013, p. 546).

The second phase is the PA phase. The aim of this phase is the evaluation and creation of a predictive model that will enable accurate prediction of phenomenon outcomes

based on a new set of observations, where ‘new’ can be interpreted as observations in future or observations that were not included in the original data sample (Kalampokis, Tambouris & Tarabanis 2013, p. 546).

The data conditioning phase consists of two steps:

1. Data collection of Twitter data
2. Data pre-processing

The PA phase consists of following steps:

3. Model evaluation and data analysis
4. Correlation and PA

Figure 1-1 summarizes the steps of each phase.

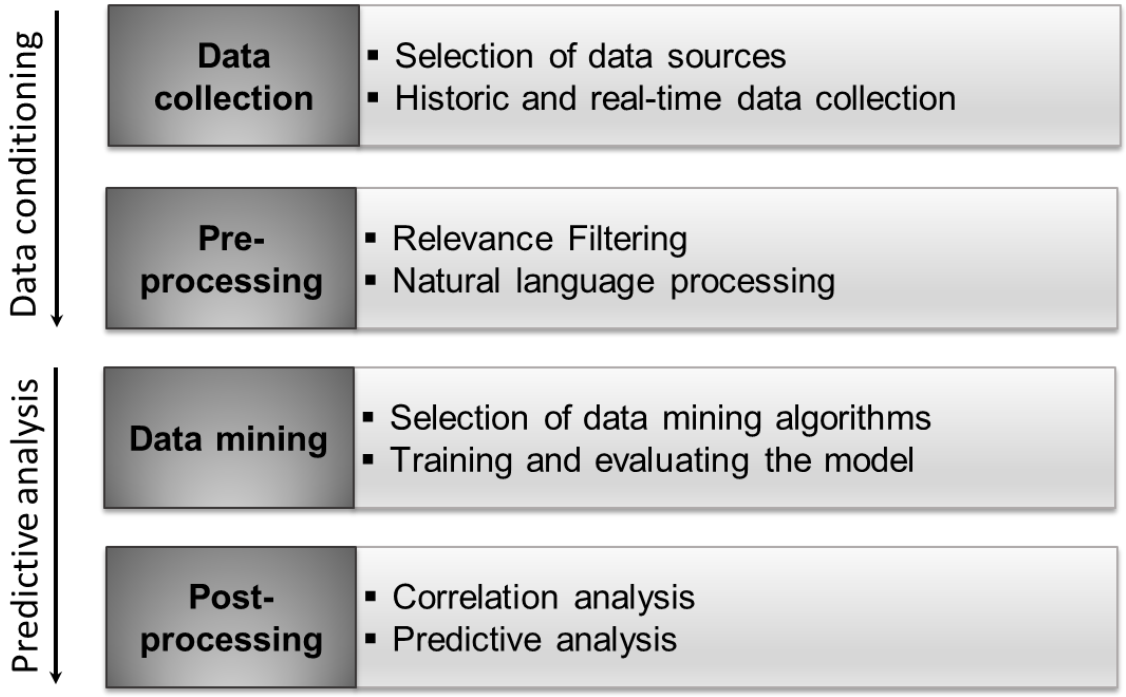


Figure 1-1: Analysis phases

It has to be noted that the steps can vary depending on the data analysis problem at hand.

1.3.1 Data conditioning phase

During the data conditioning phase, the time window has to be defined, that is, the time period over which data is collected. The query terms used for the data search have to be selected, the prediction variables, that is, the observation points in the data, have to be defined and the method for data extraction has to be evaluated. For this study, two months of tweets about Apple and its most popular products were collected. In the case of Twitter tweets, tweets can be collected either through Twitter's Application Programming Interface (API) or using a screen scraper. Other possibilities include online data collectors or data brokers. For this study the data was collected through the Twitter Search API and using the NVivo NCapture screen scraper. The Twitter Search API has the advantage that tweets can be collected automatically and the collection task can be run over an extended period of time. The prediction variables selected for this study are words in the tweets, word frequencies (bag of words approach), tweet frequencies, tweet and tokens. The collected, raw data has to be relevance filtered to remove irrelevant data such as spam or off topic tweets. Depending on the data, other tasks such as data deduplication, retweets removal or record linkage might be necessary. For this study, basic subjectivity analysis and data deduplication were performed to "clean" the data.

1.3.2 Predictive analysis phase

After the data purification steps, the model for data analysis has to be selected. Model selection happens during the PA phase. The predictive model measures the observation points. Observation points are also called the predictor variables or independent variables. Measurement of predictor variables means identifying which mood dimensions will be selected. A uni-dimensional mood model makes a binary mood distinction, that is,

positive or negative. A multi-dimensional mood model captures additional mood dimensions such as excellent, good, neutral, bad, poor or very poor. Mood dimensions can be derived from the Profile of Mood States (POMS), a well-vetted psychometric instrument (Bollen, Mao & Zeng 2010, p. 3). For this study, binary mood classification was used which classifies tweets into positive and negative tweets about Apple and its products. Selection of the predictive method will mean selecting, for example, supervised, unsupervised or semi-supervised Machine Learning (ML) algorithms. Usually, different models are trained and the best performing scheme is chosen. The selected scheme is then used to classify new, unseen data, for example tweets. Classification is a common example of a process that can be undertaken using supervised ML. Supervised methods can also be used for regression. Grouping tweets into positive and negative tweets is a binary classification task. That is why, for this study, supervised methods for binary classification and for regression were evaluated. For supervised learning algorithms, a given data set is typically divided into two parts: training and testing datasets with known class labels (Gundecha & Liu 2012, p. 2). The class labels in this study are the mood states of “positive” and “negative”. These datasets are used for the selection of the evaluation method. Typical supervised learning methods are decision tree induction, *k*-Nearest Neighbour, Naïve Bayes classification, Multilayer Perceptrons and Support Vector Machines. The best performing algorithm, which is the one with the best classification performance, will then be applied to new, unseen data for classification. In this study the Naïve Bayes, Perceptron and Multilayer Perceptron, decision Trees as well as Logistic Regression (LR) classifiers were evaluated. The prediction baseline step defines how the prediction is executed, for example by finding correlations between mood state timelines and financial data over the same period over

time. Time series analysis is a popular method for finding correlations. The classified data collected for this study is represented as time series, viz. the number of positive and negative tweets over a certain period of time. The second time series is the share price over the same period of time. The two series are evaluated to determine if they can be correlated to each other. The goal is to determine if a time series A has predictive power over a time series B. If it has, it can be used to make predictions using future, unseen data. This study used the Granger causality test to correlate the Twitter mood states time series with the share price development. Figure 1-2 summarizes these steps for a Twitter tweet.

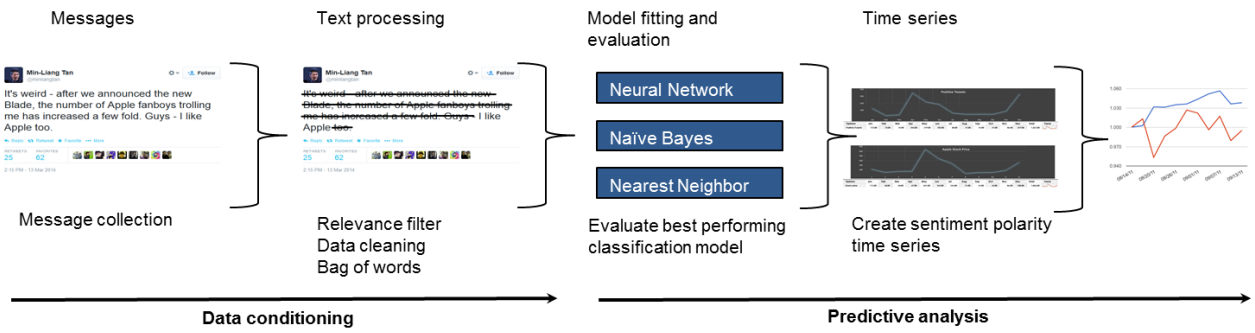


Figure 1-2: Data conditioning and pretictive analysis stages

Each step typically goes through many iterations until satisfactory results are obtained. These steps will be described in more detail in chapter 3. Research methodology.

1.4 Delimitation of the scope and key assumptions

The focus of this study is on analysing data from the Twitter microblogging website. Since microblogging messages are short and colloquial, traditional algorithms do not

perform as well as they do for long texts (Liang & Dai 2013). The SMM algorithms were evaluated in order to address the issues of scarce texts and colloquial language. However, the goal is not to find new algorithms for short text messages, but to apply suitable methods for the problem at hand.

There are many Data Mining (DM) algorithms and not all of them can be evaluated. The algorithms selected for this study are based on the ones identified in the literature review in chapter 2 Literature review. In previous studies, some algorithms have shown to perform well for similar tasks. They were evaluated and used in the framework developed for this thesis.

Free access to the Twitter search API is limited to a 15 minute quota (Twitter Developers 2015). This means that for collecting tweets, after a search query has been executed, the collection algorithm has to wait for 15 minutes before it can execute the next search query. Consequently, data collection has to run over an extended period of time. With its large and growing user base, Twitter has currently more than 317, 000, 000 monthly active users (Statista 2016) and members of different social classes are represented on Twitter. Nevertheless, there is a self-selection bias: only users who have chosen to be on Twitter are represented in the collected data. Also:

- Not everybody is using Twitter
- Not every Twitter user tweets opinions or tweets at all
- Not every tweet is giving an honest opinion

This study is not analysing the demographic distribution of its users since they can omit for instance their geographic location or age. In this case, there is no reliable way of identifying the age or location of some Twitter users.. Also, some Twitter users can pretend to be someone other than themselves or use nom de plumes, like catfish or

sockpuppet, to hide their true identity and tweet fake opinions. If these numbers are not significant, the results will only be marginally affected. However, if they are large, it will influence the outcomes. In this study only English tweets were analysed. The largest Twitter user base is in the US (Statista 2016). The assumption is that when analysing English tweets, the number is large enough that fake tweets have no significant impact on the results. Also, since a large, active user base tweets in English, the body of tweets have a representational demographic distribution in terms of age, gender, ethnic classification or social group. This study does not aim to create high quality samples but assumes that Big Data principles will smoothen uneven distributions if the datasets are large Since Big Data needs large data volumes to get meaningful results. The study focused on opinions on Apple and Apple products. Apple is a well-known company with well-established products used by many people worldwide. Hence, many people tweet about it and its products as the data collection task revealed.

1.5 Key definitions and terminologies

This chapter forms the basis for the subsequent parts of the dissertation by providing the key definitions and concepts.

1.5.1 Computational social science (CSS)

Computational social science (CSS) is the integrated, interdisciplinary investigation of social systems as information-processing organizations using the medium of advanced computational systems (Cioffi-Revilla 2010, p. 261). A CSS is emerging that leverages the capacity to collect and analyse data with an unprecedented breadth, depth and scale

(Lazer et al. 2009, p. 722). Computational social science is a fledging interdisciplinary field at the intersection of the social sciences, computational science, and complexity (Cioffi-Revilla 2010, p. 259).

CSS has two subareas relevant to the proposed research, social media mining (SMM) and sentiment analysis.

1.5.1.1 Social media mining (SMM)

Social Media Mining (SMM) is the process of representing, analysing, and extracting actionable patterns from SM data (Zafarani, Abbasi & Liu 2014, p. 2). SM data consists of the individuals, the entities such as content and sites, and the connections and interactions between the individuals. These interactions represent the social capital (SC) of SM data. SC is an investment in social relations with expected returns (Lin, Burt & Cook 2001, p. 6). SC is unlike other forms of capital in that it is not possessed by individuals, but resides in the relationships that individuals have with one another (Smith et al. 2011, p. 2). SMM takes into consideration these relations of the data when analysing it. Analysing SM data is the task of mining user-generated content with social relations (Zafarani, Abbasi & Liu 2014, p. 2).

1.5.1.2 Sentiment analysis

Sentiment analysis, also called opinion mining, is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (Liu 2012, p. 1). In SMM, sentiment analysis is then the automated extraction of emotional content from SM data (Jones & Huan 2013, p. 94).

To extract the opinions from user generated content, the natural language has to be transformed into a more formal representation such as a tree structure that is easier to interpret for a computer. This technique is called Natural Language Processing (NLP), a field in computer science and linguistics which is concerned with the interactions between computers and human (natural) languages (Olive, Christianson & McCary 2011, p. 1).

1.5.2 Big Data

A universally agreed upon definition of "Big Data" has not been developed to date (Finlay 2014). Big Data refers to data that is too big to fit on a single server, too unstructured to fit into a row-and-column database, and/or too continuously flowing to fit into a static data warehouse (Davenport 2014, p. 1). Big Data does not refer to large data volumes alone. Big Data also has increased velocity (i.e., the rate at which data is transmitted and received), complexity, and variety compared to data sources of the past (Franks 2012, p. 5). This is usually referred to by the three V's: Volume, Velocity and Variety. Big Data techniques are used to handle large-scale datasets, find useful patterns and gain insights and knowledge. Big Data analysis is a process of knowledge discovery from raw data.SM data are largely user-generated content on SM sites (Gundecha & Liu 2012, p. 4). The pervasive use of SM has generated unprecedented amounts of social data (Gundecha & Liu 2012, p. 1). SM data can be considered Big Data due to the amount and the speed at which the data is generated. SM data are vast, noisy, unstructured, and dynamic in nature, and thus novel challenges arise (Gundecha & Liu 2012, p. 1). As such, Big Data has two subcategories, data mining (DM) and predictive analysis (PA).

1.5.2.1 Data mining (DM)

Data mining (DM), also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories, or data streams (Han, Kamber & Pei 2011, p. xxiii). DM involves statistical and/or artificial intelligence analysis, usually applied to large-scale datasets (Olson & Delen 2008, p. 4).

Throughout the literature review no consensus was found to distinguish Big Data and DM. Some authors define DM as the analysis of large and complex datasets (Finlay 2014). Others define DM as the process of information generalization (Menasalvas & Wasilewska 2006). Some authors define DM as the analysis of the datasets themselves and Big Data the analysis of the relations between the datasets. In this thesis the terms are used interchangeably. SMM is a form of Big Data analysis based on SM data. It finds patterns and relations in SM data that can be used for better understanding customer needs, voters' preferences or detecting future crisis. DM turns a large collection of data into knowledge (Han, Kamber & Pei 2011, p. 2).

1.5.2.2 Predictive analytics (PA)

The research for the thesis is in the area of predictive analysis (PA). PA uses advanced analytics for predictive modelling. Advanced analytics goes further than core analytics. Advanced analytics includes everything from complex ad hoc SQL, to forecasting, to DM, to predictive modelling (Franks 2012, p. 187). PA refers to "Technology that learns from

experience (data) to predict the future behaviour of individuals in order to drive better decisions" (Siegel 2013, p. 11).

1.5.3 Artificial intelligence (AI)

Artificial intelligence (AI) is a branch of computer science. The central scientific goal of AI is to understand the principles that make intelligent behaviour possible in natural or artificial systems (Poole & Mackworth 2010, p. 4). In the literature there are many ways AI is defined. AI deals with symbolic, non-algorithmic methods of problem solving. In this thesis, AI is referred to as the study of making computers do things associated with tasks humans do better at the moment. AI is that part of computer science concerned with designing intelligent computer systems that exhibit the characteristics associated with intelligence in human behaviour (Akerkar 2005, p. 2). ML is an area of AI where computers have the capability to learn. Learning involves an agent remembering the past in a way that is useful for the future (Poole & Mackworth 2010, p. 283). ML techniques are widely used in DM since they can generate rules that would be too complex or too many for a developer to program manually.

1.5.3.1 Machine learning (ML)

Machine learning (ML) is the study of data-driven methods capable of mimicking, understanding and aiding human and biological information processing tasks (Barber 2012, p. xv). ML is also closely aligned with AI, with ML placing more emphasis on using data to drive and adapt the model (Barber 2012, p. xv). Learning is the ability of an agent to improve its behaviour based on experience (Poole & Mackworth 2010, p. 283). Since computers do not have experiences, they learn from data. Contrary to the data

conditioning steps, ML algorithms are not domain specific and can be used for any data mining task. ML algorithms have been applied to many natural language processing problems, including text classification, part of speech tagging, parsing, named entity recognition, word sense disambiguation, etc., each requiring a large number of labelled examples such as documents and their classes, words in the context and their senses as training data (Law & Ahn 2011, p. 9).

The application of ML methods to large databases is called DM (Alpaydin 2004, p. 2). These datasets are often too large for traditional DM techniques. Also, ML techniques have been used in NLP because understanding and interpreting natural language has been a challenging task with conventional techniques such as lexicon-based methods (Liu 2012, p. 119). ML schemes can be trained with a large corpus of text documents to mitigate the short comings of traditional methods. As such, the building of machines to automatically parse and understand natural languages has been a central endeavour for researchers in Artificial Intelligence (AI), Information Retrieval (IR) and Natural Language Processing (NLP) (Law & Ahn 2011, p. 9). However, it is also useful to realize that sentiment analysis is a highly restricted NLP problem because the system does not need to fully understand the semantics of each sentence or document but only needs to understand some aspects of it, i.e., positive or negative sentiments, and their target entities or topics (Liu 2012, p. 13). Also, opinion mining is highly domain specific. Training data for one domain might perform poorly on test data from another domain.

1.5.3.2 Data classification

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The models are derived based on the analysis of

a set of training data (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown (Han, Kamber & Pei 2011, p. 18). In their work Han et al (2011) focus on the feasibility, usefulness, effectiveness, and scalability of techniques of large datasets. In this study, classification techniques were used to classify tweets into their sentiment polarity over a certain time frame.

1.5.4 Theoretical framework

The theoretical framework is based on AI, CSS and Big Data. Figure 1-3 shows the research question in its theoretical framework:

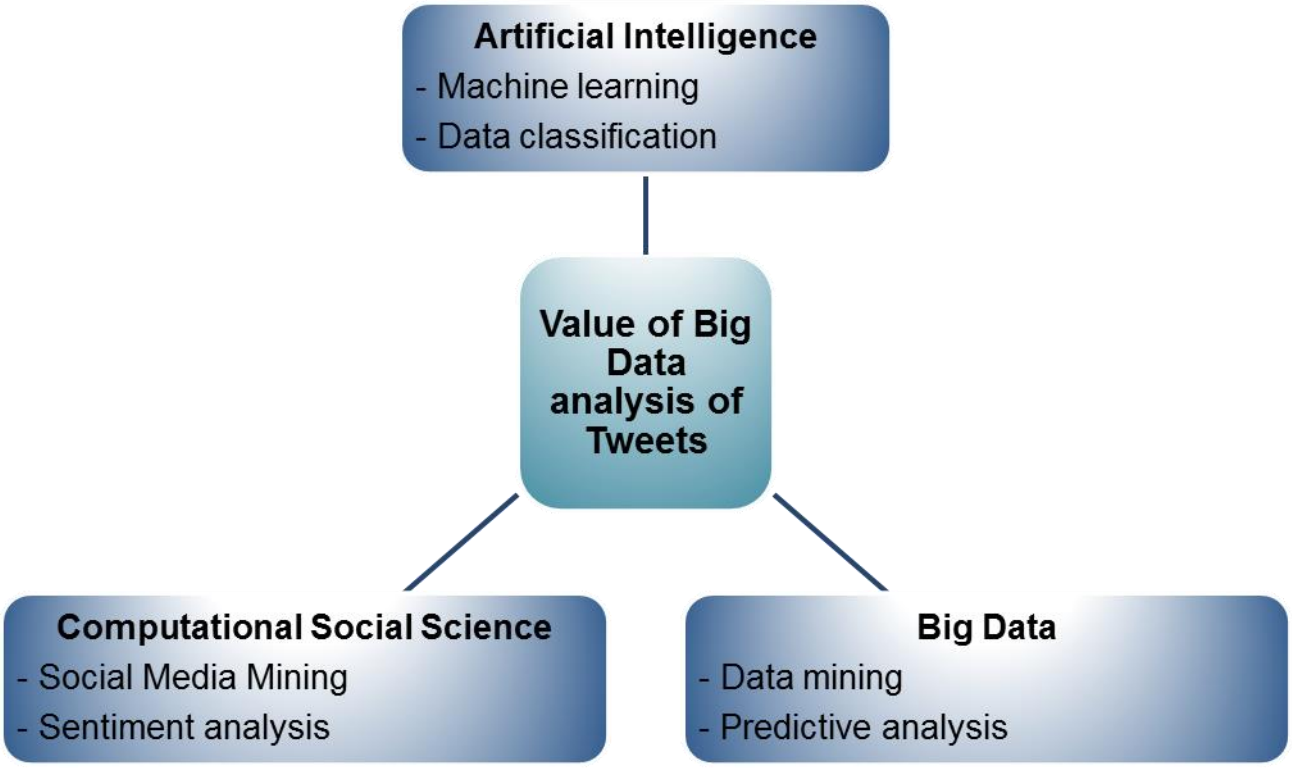


Figure 1-3: Theoretical framework

A framework developed for this study was used for analysing Twitter tweets and finding correlations with share prices. It used DM techniques for collecting and analysing data and ML techniques for data classification and finding correlations. Its contribution is to determine how SM mining, viz. Twitter tweets, can be used for PA for financial markets. Several SMM, sentiment analysis, PA and Big Data techniques were evaluated to build the framework. The best performing ones were then used on the actual data to be classified.

1.6 Publication list

As part of this thesis, several papers and book chapters were published, and some of the research was presented at conferences. Table 1-1: Publication list lists the papers that were published as part of the study:

Authors	Title	Publication	Year	Publisher
Wlodarczak, Peter	Smart Cities, Enabling technologies for future living	City Networks - Planning for Health and Sustainability	2017	Springer International Publishing
Wlodarczak, Peter	Cyber Immunity, A Bio-Inspired Cyber Defense System	International Conference on Bioinformatics and Biomedical Engineering	2017	Springer International Publishing
Wlodarczak, Peter; Soar, Jeffrey; Ally, Mustafa	Context Aware Computing for Ambient Assisted Living	International Conference on Smart Homes and Health Telematics	2016	Springer International Publishing
Wlodarczak, Peter; Soar, Jeffrey; Ally, Mustafa	Genome mining using machine learning techniques	International Conference on Smart Homes and Health Telematics	2015	Springer International Publishing
Wlodarczak, Peter; Soar, Jeffrey; Ally, Mustafa	Multimedia data mining using deep learning	IEEE Xplore	2015	IEEE
Wlodarczak, Peter; Qian, Siyu; Ally, Mustafa; Soar, Jeffrey	Social genome mining for crisis prediction	21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining	2015	PopInfo'15
Wlodarczak, P.; Soar, J.; Ally, M.	Behavioural health analytics using mobile phones	EAI Endorsed Transactions on Scalable Information Systems	2015	European Union Digital Library

Wlodarczak, Peter; Soar, Jeffrey; Ally, Mustafa	Reality mining in eHealth	International Conference on Health Information Science	2015	Springer International Publishing
Wlodarczak, Peter; Soar, Jeffrey; Ally, Mustafa	Big Data Analysis, from Cloud to Crowd	SSRN	2015	Elsevier
Wlodarczak, Peter; Soar, Jeffrey; Ally, Mustafa	Big Data analytics of Social Media	19th International Conference on Circuits, Systems, Communications and Computers, (CSCC 2015)	2015	INASE
Wlodarczak, P; Soar, J; Ally, M	What the Future Holds for Social Media Data Analysis		2015	World Academy of Science, Engineering and Technology
Wlodarczak, Peter; Ally, Mustafa; Soar, Jeffrey	Data Process and Analysis Technologies of Big Data	Networking for Big Data	2015	Chapman and Hall/CRC
Wlodarczak, Peter; Ally, Mustafa; Soar, Jeffrey	Opinion Mining in Social Big Data	SSRN	2015	Elsevier
Wlodarczak, Peter	An approach for big data technologies in social media mining	Journal of Art Media and Technology	2015	JAMT
Wlodarczak, Peter	Big Personal Data	SSRN	2014	Elsevier

Table 1-1: Publication list

At the time of writing, the book chapter “Smart Cities, Enabling technologies for future living” had not been published yet.

1.7 Structure of the thesis

The first chapter outlines the study areas and defines the goals and aims of the study. The research question is formulated and the underlying hypotheses are stated. An overview of the methodology is presented and the scope is defined along with the key definitions and terminology.

The second chapter reviews the literature to build the theoretical foundation of the study and gives an overview of research related to the area of this thesis that has been conducted thus far. It identifies key issues which form the basis of the research.

The third chapter describes the theoretical framework of the research and refines the research question and issues.

The fourth chapter provides a detailed description of the adopted research methodology. It describes the data collection and analysis methods and the approach taken to find correlations. Finally, the limitations and ethical concerns are addressed.

The fifth chapter presents the results and findings. It sets out summary tables and figures of the results to enhance readability. Chapter 5, Conclusions and implications, presents the conclusions and implications for theory, practice and methodology. The limitations that came up during the research and suggestion for future research are elaborated.

2 Literature review

2.1 Introduction

The previous chapter identified the research problem and provided an overview of the research methodology. This chapter reviews the current literature upon which the theoretical basis of this thesis is built. The two main areas of literature reviewed were in the areas of CSS and Big Data. Based on the interdisciplinary nature of the research there are several related areas covered in this review. Also, there are overlaps, for instance there is no clear separation between CSS and Big Data. SMM is a form of Big Data analysis and here also is no clear distinction between Big Data analysis and DM. Different authors use these terms interchangeably (Alpaydin 2004; Menasalvas & Wasilewska 2006; Finlay 2014). The review focuses on literature directly relevant to the research. ML and classification are covered in the sections on literature on SMM and PA since they are supporting theories in these fields.

A considerable amount of research has been performed on using Twitter data (Arias, Arratia & Xuriguera 2014; Paltoglou & Thelwall 2012; Achrekar et al. 2011). Twitter has several characteristics that make it a popular research subject. Firstly, unlike in other SM sites such as Facebook or Google+ where access to posts can be limited to friends, all tweets are by definition public. Secondly, tweets are limited to 140 characters which simplifies NLP. Generally, no complicated linguistic constructs are used in posts because of the character limitation. Thirdly, Twitter provides a powerful search API to access historical data. When using appropriate queries much of the data pre-processing, such as relevance filtering, can already be handled in the data collection step.

The literature review is based mostly on searches in the University of Southern Queensland (USQ) online library and on Google Scholar, but also on other online libraries and indexing sites such as CiteSeerX or arXiv.org.

The search terms were initially selected to focus on the research areas as described in chapter 1.5.4 Theoretical framework. Then they were narrowed down to include only search results in the specific area, for instance PA, PA using SM, PA using Twitter. Table 2-1: Literature search terms lists the search terms that were used:

Data Mining
Big Data
Artificial Intelligence
Machine learning
Predictive Analytics
Opinion mining
Data mining
Social media mining
Social media analysis
Twitter analysis
Predictive analysis using social media
Predictive analysis using Twitter
Big Data analysis using social media
natural language processing
social media using machine learning
behavioural analytics
natural language processing using social media
opinion mining on social media
opinion mining for predictive analytics
research methodology
research framework
theoretical framework
predicting share price
predicting financial indicators

Table 2-1: Literature search terms

2.1.1 Computational social science

There is an extensive literature about CSS. The literature review focused on literature relevant for this thesis which is CSS itself and opinion mining.

2.1.1.1Social media mining

SM data is big, linked, noisy, highly unstructured and often incomplete. Therefore it differs from data in traditional DM, which fosters a new research field - SMM (Tang, Chang & Liu 2014, p. 20). In recent years, a great deal of research has gone into analysing SM, and a large body of new studies has been published (Xinyu, Youngwoon & Suk Young 2015; Wei, Mao & Wang 2015; Tsakalidis et al. 2015). SMM is an interdisciplinary field rooted in computer science and social sciences. Researchers in this emerging field are expected to possess knowledge in different areas, such as DM, ML, text mining, social network analysis, and information retrieval. They are often required to consult research papers to learn the state of the art of SMM (Zafarani, Abbasi & Liu 2014, p. 10). Among other DM algorithms, ML techniques have been used for SMM (Souza et al. 2015; Wlodarczak, Soar & Ally 2015; Shulong et al. 2014; Liu et al. 2014; Gerber 2014; Li et al. 2014). ML techniques are well suited when a problem cannot be adequately solved using simple (deterministic), rule-based solutions, when the rules are too complex or when you cannot scale. ML algorithms are trained with a set of data to recognize, for example the mood polarity (sentiment) of the data. ML is an area of AI and is not a recent technology. ML goes back to the late sixties when Arthur Samuel defined ML as the ability of a computer to learn without being explicitly programmed (Samuel 1959). However, due to the nature of SM data, and for many Big Data problems, they

provide very suitable techniques that can handle the large volumes of data and their heterogeneous nature. Since the invention of the first ML algorithms in 1959 (Samuel 1959), many new ML algorithms have been developed and are widely used now for SMM and Big Data analysis.

DM “algorithms cover classification, clustering, statistical learning, association analysis, and link mining, which are all among the most important topics in DM research and development” (Wu et al., 2007, p. 2). Wu et al. (2007) have identified the top ten algorithms used in SMM. These algorithms are at the heart of SMM. They identified the following algorithms, C4.5, k-Means, SVM (Support Vector Machines), Apriori, EM (Expectation Maximisation), PageRank, AdaBoost, kNN (k-Nearest Neighbor), Naïve Bayes, and CART (Classification And Regression Trees), which are considered to be among the best techniques in this area. Most belong to the family of ML algorithms except for PageRank, which is based on webgraphs and graph theory. Wu et al. (2007) conclude that they are “among the most influential algorithms for classification, clustering, statistical learning, association analysis and link mining” (Wu et al., 2007, p. 34).

AdaBoost belongs to a family of learners called ensemble learners. If several learning schemes are available, it may be advantageous not to choose the best-performing one for your dataset (using cross-validation) but to use them all and combine the results (Witten, Frank & Hall 2011, p. 351). Ensemble learners have been used in many studies with surprisingly good results; such as for genome-wide prediction of traits (González-Recio, Rosa & Gianola 2014), for approximating the crowd to predict majority opinions (Ertekin, Rudin & Hirsh 2014), for SMM (Tang, Chang & Liu 2014), for stock market prediction (Bouktif & Awad 2013), and for decision support systems (Pardeep et al.

2012). Ensemble learners work well in noisy environments, which is a typical characteristic of SM data. Even adding noise can improve the results in some cases. Noise can be described as items that carry no content of knowledge (Du 2013, p. 63). AdaBoost is an ensemble learner that has been developed specifically for classification. It can be applied to any classification learning algorithm (Witten, Frank & Hall 2011, p. 359). Unfortunately, ensemble classifier methods do not take into account the interpretability of final classification (Bouktif & Awad 2013, p. 837). This lack of interpretability makes it difficult to determine which factors contribute to the result and to what extent.

To retrieve knowledge from SM data, the data has to be analysed semantically. Many studies covered the techniques of mining text for opinion and sentiment analysis (Petz et al. 2014; Huang et al. 2013; Kao et al. 2013; Paltoglou & Thelwall 2012; Zeng et al. 2010). Sentiment analysis is a subarea of Natural Language Processing (NLP). There are many NLP techniques such as stemming, lemmatization, or part-of-speech tagging. A common approach is splitting sentences into n-grams (Zlacky et al. 2014; Oliveira, Cortez & Areal 2013; Lloret et al. 2012), where n can be one for only one word, a unigram, a bigram for two words, trigrams etc. A different approach was pursued by Neri et al. (2012). They performed a sentiment analysis study on Facebook users comparing their sentiment towards RAI, the public Italian broadcast company against the private company LA7. They did not only use the positive or negative polarity of unigrams but instead analysed the whole sentences on the syntactic tree. They concluded that customer monitoring is a good way to measure loyalty and keep track of their sentiment towards brands and products. Whole sentence analysis has the advantage that it takes into account the context of sentiment words.

A different approach in the area of SMM is also used by Schreck & Keim (2013). They use a visualization approach to perceive information in SM. They point out the advantages of visual analysis to extract multifaceted information and correlate it with textual, geospatial, temporal, and other contextual data. This approach can be used on such data to track opinions about new products and services, fads and trends in popular culture, adverse reactions to prescription drugs, infectious disease epidemiology, fraud and other types of criminal activity, the public's response to a political candidate or proposed legislation, motor vehicle defects, and different groups' consumption habits (Schreck & Keim 2013, p. 69). However, they do not describe the details of their methodology or techniques that they used. SMM has become mainstream in research and new studies continue to appear analysing SM from different perspectives Xinyu, Youngwoon & Suk Young (2015) analysed tweets to predict the time and location where a specific type of crime was likely to occur. They used sentiment analysis of tweets and correlated it with weather data and historic crime data from a database. By correlating sentiment polarity and external influences, in their case weather data, they obtained better results from a hot-spot kernel density estimation model on theft incidents. Adding external influences to the analyses used in studies such as marketing campaigns or product announcements could be potential fields of future research. Wei, Mao & Wang (2015)) investigated the relationship between Twitter volume spikes and stock options pricing. They concluded that stock volatility around a Twitter volume spike and found that a three-parameter model that used the same drift and different volatilities before and after a Twitter volume spike provided the highest gain in the likelihood value. The work of Tsakalidis et al. (2015) focused on exploiting Twitter's content to predict the 2014 EU election results in Germany, the Netherlands, and Greece. They used users' voting

intentions and treated it as time-variant features. They employed time series analysis as well as sentiment analysis of tweets. They did not train a classifier but applied a lexicon-based approach for sentiment analysis. Ishijima, Kazumi & Maeda (2015) analysed the sentiment towards the Japanese economy that might appear in daily news articles. They used word frequencies to classify articles into positive or negative articles about the current economic situation in Japan. They then constructed a daily summary index and performed statistical analysis to examine correlations between the sentiment index and Tokyo Stock Exchange prices. They concluded that the index significantly predicts stock prices of three days in advance. This study also used frequency analysis; however the prediction accuracy was lower from more sophisticated classifiers.

2.1.1.2 Sentiment analysis (SA)

Sentiment analysis (SA) using SM data has become a very active area of research. SA, also called opinion mining, on SM has not only attracted interest from academia, but also from the industry. For the first time in human history, we now have a huge volume of opinionated data in SM on the Web (Liu 2012, p. 8). In the past, companies have had to conduct surveys or opinion polls to obtain consumer opinions on their products and services. Using SA, it is possible to obtain the opinions of millions of users from SM data. Liu (2012) analyses the semantic orientation of an opinion using a feature-based opinion mining model where a feature is a finite set of words or phrases. He then tries to discover the hidden pieces of information by comparing an evaluative document against the comments of an opinion holder. Tsvetovat, Kazil and Kouznetsov (2013) focus on SA in social networks. They describe the challenges of SA as a replacement for polling, and

developed a new approach they called Implicit Sentiment Mining to overcome some of the difficulties. They rely on a psychological phenomenon called mirroring. To use mirroring, they collected a sample of texts from speeches of political candidates and Twitter posts from the same period which they processed through a linguistic pipeline to gauge overall party preference and propensity to vote for a specific candidate. Pang and Lee (2008) analysed methods for opinion mining and SA. They asserted that opinions are an important factor in the decision making process. They proposed a search specialized engine to get opinions and applications for different kinds of opinion Web sites. They provided, among others, an application for Business Intelligence. Taking this a step further, our proposed study quantifies the value of such an application. They provide a detailed description on how opinions were extracted, classified and summarized, both text based and graphic based. They also cover credibility, net authority and net influence.

2.1.1.3 Quantifying social media data

There have been many attempts to capitalize on the so-called “wisdom of the crowd” (Schoen et al. 2013). However, quantifying SM data remains challenging since it is often difficult to determine the impact of SM posts on human behaviour (Probst, Grosswiele & Pfleger 2013). Several studies tried to quantify SM data, many focusing on targeted marketing and on the return on investment (ROI) from marketing campaigns on SM (Zhang 2013; Goh, Heng & Lin 2012; Gomez-Arias & Genin 2009; Clemons 2009). Advances in information technology, data gathering and analytics are enabling companies to manage all phases of the customer life cycle, including acquiring new

customers, increasing revenue from existing customers and retaining new customers (Garcia Martinez & Walton 2014). As traditional advertising is losing its impact, both advertisers and the media owners who are dependent upon them are desperately seeking alternative ways to reach consumers and alternative ways to earn revenue (Clemons 2009, p. 46). Goh, Heng and Lin (2012) analysed the effect of direct marketer efforts on a Facebook fan site of an undisclosed company. They used quantitative approaches based on Heckman selection models and used control groups who have no Facebook account. They concluded that, ceteris paribus, the group that was targeted by direct marketing has a higher propensity to buy than an untreated group. They used econometrics to quantify the economic value for marketing efforts. The authors provide suggestions for future research but do not give clear indications of how to apply their findings in practice for online marketing efforts.

Mayer (2009) analysed how socioeconomic background and incentives affect the structure and composition of social networks. He reviewed the theoretical and empirical literature studying these relations and possible implications of internet based social interactions. He concentrated on who interacts with whom, and how and why networks are formed. He analysed the influence of social networks on decision-making, trade, education, labour markets. He found that online social networks are formed in similar ways as offline social networks come about. However, the communication patterns differ in online social networks. He concluded that online social networks improve information transmission, foster price transparency, facilitate learning and make it easier to obtain product features or characteristics of trade partners. There was no indication however on what the implications of the study for policy and practice are or what the concrete applications are.

2.1.1.4 Frameworks for computational analysis

Several attempts have been made to develop a framework for SM analysis (Arias, Arratia & Xuriguera 2014; Nguyen, Yan & Thai 2013; Nohuddin et al. 2012). They typically consist of a data collection and pre-processing mechanism, a data analysis engine and a visualisation and reporting frontend. Depending on the purpose of the framework, using correlation analysis (Souza et al. 2015; Siganos, Vagenas-Nanos & Verwijmeren 2014; Arias, Arratia & Xuriguera 2014; Kalampokis, Tambouris & Tarabanis 2013; Bollen, Mao & Zeng 2010) or predictive analysis (Burnap et al. 2015; Ishijima, Kazumi & Maeda 2015; Liu et al. 2014) is part of the framework. McKelvey et al. (2012) developed a framework they call Truthy, a system for collecting and analysing Twitter data of political discourses. Truthy is a system for visualizing political information diffusion on Twitter. Truthy can render statistical and visual overviews of large-scale communication networks. Truthy treats every tweet to be a meme, and categorises and visualises them as themes. Using this approach, they try to bridge the quantitative and qualitative epistemologies. They are interested in the visualisation aspect and not in opinion mining. They do not give insights into the metrics used.

Xu, Li & Song (2012) propose a framework to identify the most valuable customers to maximize the profit of an enterprise. They conducted an empirical study and used an optimization technique based on semidefinite programming. They showed how, based on online influence and authority, more targeted marketing and reputation management can be performed.

Smith et al. (2011) have tried to quantify social capital in online communities. They developed a computational framework to analyse cross-references, bonding, affinities and other kinds of relationships. They incorporated resources such as jobs, moral

support etc. more fully into their social capital framework from previous studies with a limited number of participants (Smith et al. 2011, p. 3) and showed how these resources can be mobilized. They conducted a proof of concept with four case studies. However, there is no single value for social capital. Also there is no economic quantification given, for example finding a job through a social network compared to finding one through a head hunter.

Chau and Xu (2012) developed a framework to analyse business intelligence on blogs. Their primary focus was to analyse the blog content and the underlying social networks. The main purpose of the framework is to shape online reputation and improve marketing efforts. As proof of concept, they provided two case studies, one on a company, Starbucks, and one on a product, the iPod. They clearly describe their research design, detailing how they do their business intelligence collection and data analysis. The case studies described step-by-step how the material was gathered and analysed. However they did not mention how the automated steps were executed or what tools they used. Also, they did not describe the concrete value of their framework to BI.

The research papers described in the section on Frameworks for computational analysis provide the starting point for the proposed study because while they had developed frameworks for a different purpose, they employed theories and techniques that have the potential for use in our study.

Table 2-2 summarizes the literature review described in section 1.5.1

Challenge	Source	Synopsis
Social Media mining	Souza et al. (2015); Shulong et al. (2014); Liu et al. (2014); Gerber (2014); Li et al. (2014) ; Wu	Techniques for analysing structured, semi-structured and unstructured data are discussed. ML techniques

	et al. (2007) ; Klein, Tran-Gia & Hartmann (2013); Berman (2013); Mayer-Schonberger & Cukier (2013); Zafarani, Abbasi & Liu (2014); Wlodarczak, Soar & Ally (2015);	for classification problems are covered and under which conditions they are used. ML techniques have been widely applied to SM mining problems and provided good results since they can handle the noisiness and heterogeneity of SM data. They identified following algorithms C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, k-NN, Naive Bayes, and CART as the top of breed. They conclude that they are “among the most influential algorithms for classification, clustering, statistical learning, association analysis and link mining” (Wu et al., 2007, p. 34).
Sentiment analysis	Paltoglou & Thelwall (2012); Huang et al. (2013); Kao et al. (2013); Zeng et al. (2010); Liu (2012); Tsvetovat, Kazil & Kouznetsov (2013); Pang and Lee (2008)	SA techniques are discussed, how they can be applied to SM and how they influence decision making. Text mining and sentiment analysis were used in SM analysis to determine customer loyalty. Sentiment analysis was also used to predict elections or overall party preference during elections and customer behaviour. The studies concluded that accurate predictions could be made when adopting appropriate psychological

		and linguistic techniques.
Quantifying social media data	Goh, Heng and Lin (2012); Mayer (2009); Manyika et al. (2011); Zhang (2013); Goh, Heng & Lin (2012); Gomez-Arias & Genin (2009); Clemons 2009	SM data has been quantified in terms of marketing efforts. The influence on decision making, trade, education and labour markets has been investigated. SM data can improve marketing efforts, especially for personalised and directed marketing campaigns. They conclude that Big Data can create significant value to the economy. They give estimates on how much money can be saved for example in the health sector if it were to use Big Data creatively. The studies conclude that determining the value is still challenging.
Frameworks for computational analysis	McKelvey et al. (2012); Smith et al. (2011); Chau and Xu (2012)	SM analysis frameworks have been developed for analysing political discourses, determining the social capital in online communities and to analyse Business Intelligence in blogs. They propose to enhance the frameworks for more general purpose SM analysis.

Table 2-2: Computational social science literature overview

2.1.2 Big Data

In a broad range of application areas, data is being collected at an unprecedented scale (Jagadish et al. 2014, p. 86). The volumes of the collected data are often too large to be processed by traditional database technologies. Big Data has provided the technologies to store and analyse these large scale datasets. However, in literature there is no consensus on the definition of Big Data. Some authors use Big Data to refer to the data itself (Twinkle & Paul 2014; Berman 2013), others denote the technologies to process it (Mayer-Schonberger & Cukier 2013). Here the term Big Data is used to refer to the data analysis methods used for Big Data.

Big Data has some fundamental differences when it comes to data collection and analysis. For one, no sample selection is needed since all or almost all data can be analysed (Mayer-Schonberger & Cukier 2013), for instance all opinions on Twitter. Big Data does not necessarily mean a huge quantity of data. For instance, 100,000 records are a small quantity for a database to manage but to perform a survey among 100,000 people is an impossible task for a human with the result that not all people would be included in the survey but only a sample will be selected. This is not necessary in Big Data analysis. In this sense, Big Data means “all” or almost all data, for example, all people who retweeted a tweet.

Secondly, in a world of small data (Berman 2011, p. 154) the sample has to be of high quality. “Since we only collected a little amount of information, we tried to make certain that the figures we bothered to record were as accurate as possible” (Mayer-Schonberger & Cukier 2013, p. 32). This is not necessary for Big Data. For instance, if some subjects included in the analysis do not actually give their opinion on the iPhone

but some other product, the result would not necessarily influence the outcome because of the large quantity of opinions.

Big Data is typically unstructured and messy. Unlike in a relational database, where data is organised in tables and containers, SM data is an accumulation of text, binary data such as multimedia data, or both such as text data with embedded pictures or videos. The data is unstructured and categorised by the users. For instance, tweets do not have a subject field. When a user writes a tweet, he or she usually “tags” it using the hash “#” sign. In other words, data is being labelled. Tagging has become the de facto standard today for categorising content on the internet. There is no standardised way for tagging. Some tags might be misleading or simply wrong. Nevertheless, noisy data does not falsify the validity of the result if there is enough of it. In Big Data analysis, more trumps better. Big Data transforms figures into something more probabilistic than precise (Mayer-Schonberger & Cukier 2013, p. 35).

It is important to mention that Big Data finds correlations between data, not causes. For instance, Google publishes influenza trends based on flu related searches and medical records in a certain region (Ginsberg et al. 2009). If there are many searches it assumes there is a flu epidemic in that area. However, it does not find the cause of the epidemic. Big Data analysis finds the what, not the why. Accordingly, this study focuses on finding correlation, not causation.

Big Data analysis now drives nearly every aspect of society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences (Jagadish et al. 2014, p. 86). The literature on Big Data for this study falls into two categories, DM and PA. Many studies cover techniques and methods of Big Data analysis (Wright 2014; Dinu & Iovan 2014; Klein, Tran-Gia & Hartmann 2013; Berman 2013). These techniques

include distributed computing (Twinkle & Paul 2014), real-time analytical processing (Cheng-Zhang et al. 2014), distributed storage (Shvachko et al. 2010) and Big Data architectures (Chan 2013). Big Data technologies are a separate area of research and are beyond the scope of this thesis.

Some studies also cover the limitations of Big Data. Buhl et al. (2013) argue that there have to be novel, innovative business models to fully exploit its usefulness. Some authors also cover the anonymity issues arising with Big Data analytics (Tucker 2014; Boyd & Crawford 2012) and privacy-preserving DM has been extensively discussed in the literature to overcome this problem (Agrawal & Srikant 2000).

Manyika et al. (2011) tried to determine how Big Data could reduce costs in different industries. They concluded that Big Data can create significant value to the economy. It can enhance productivity and efficiency in the private and public sectors. They suggest estimates on how money can be saved for example in the health sector if it were to use Big Data creatively. Their study identified the areas where and how Big Data can help improve decision making, discover needs or create greater transparency. The conclusion was that Big Data will change the way business is done today and will increase the competitiveness, productivity and efficiency of companies. However they do not specify how they came up with their estimates and what methods they used to arrive at them.

2.1.2.1 Data mining (DM)

Utilizing appropriate data mining (DM) algorithms is crucial to obtaining meaningful results. The algorithms have to be efficient and suitable for a Big Data setting. Many studies have developed and analysed DM algorithms (Lim, Chen & Chen 2013;

Bulysheva & Bulyshev 2012; Kumar et al. 2012; Zeng, Li & Duan 2012; Trif 2011). These algorithms are the basic building blocks for the framework developed for our study.

Trif (2011) compared classic back-propagation to genetic algorithm based training in mobile systems based on performance and resource consumption. Back-propagation is used for training neural networks. After every training iteration, the error is propagated back through the network, and the weights of the axon connecting the neurons are adapted. Genetic algorithms are inspired by evolution and are used for optimization problems. After each generation the best results are selected and their best features are employed in the next generation. Their author concluded that genetic algorithms are more efficient especially in the mobile context and from a security perspective since they use fewer resources from traditional back propagation algorithms. Unfortunately, The study does not mention the nature of the data nor the quantity. It would be expected that this would influence the efficiency greatly. Kumar et al. (2012) analysed an evolutionary approach based on hybrid classification models in datasets from different domains. The classification of data is one of the main tasks in DM, but selecting the appropriate model to use depends on a number of different factors. Various issues such as predictive accuracy, training time to build the model, robustness and scalability need to be considered and can have trade-offs, further complicating the quest for an overall superior method (Kumar et al., p. 25). They concluded that the use of the genetic algorithm is the most consistent algorithm in terms of predictive accuracy and the use of decision trees in terms of training time. A genetic algorithm is the first choice when predictive accuracy and comprehensibility are the selection criterion and decision tree is the first choice when training time is a selection criterion (Kumar et al., p. 40). However, there is no mention of the application of the outcome to real world problems.

Lim, Chen & Chen (2013) analysed current Business Intelligence (BI) research in the areas of Big Data, text analytics and network analytics. They reviewed state-of-the-art techniques and algorithms and determined directions for future research. They found that a great amount of research is currently being undertaken in the areas of both descriptive and predictive models in relation to user preferences and behaviour (Lim, Chen & Chen 2013, p. 9). This paper was a starting point for several technologies and techniques for this thesis. Bulysheva & Bulyshev (2012) proposed a new segmentation algorithm for identifying data that is valuable for the business. They presented a new algorithm for data segmentation to help build time-based customer behaviour models. The proposed study will build on this and determine the economic value of such behaviour models. Zeng, Li & Duan (2012) reviewed current Business Intelligence (BI) systems with an emphasis on BI algorithms. They were critical of current BI systems that only analysed past data and did not attempt to conduct any projections.

Akhtar, Zamani and El-Sayed (2012) tried to find associations between individual data records or datasets. They modified some a priori and frequency algorithms to find Boolean associations in a BI system. This technique was used in a case study on al medical database as proof of concept. The conclusion was that the modified algorithm improved time and space complexity in association finding. Association finding is an important process in the proposed study since it can determine net authority and influence and thus identify important associations from less important ones. Their algorithms could turn out to be useful for our study because it provides improved association algorithms.

2.1.2.2 Predictive analysis (PA)

Early research on stock market prediction was based on random walk theory and the Efficient Market Hypothesis (EMH) (Bollen, Mao & Zeng 2010, p. 1). However, stock market prices do not follow random walks and EMH is unreliable since financial news are unpredictable. New approaches in predictive analysis (PA) of financial markets have used SM data, often using SA and ML approaches.

Twitter analysis has been used for predicting stock market development (Wei, Mao, & Wang 2015; Souza et al. 2015; Bollen, Mao & Zeng 2010), election outcomes (Burnap et al. 2015; Tsakalidis et al. 2015; Arias, Arratia & Xuriguera 2014; Tumasjan et al. 2011) and box office sales (Liu et al. 2014; Wong, Sen & Chiang 2012; Asur & Huberman 2010).

One of the first studies trying to link Twitter moods to financial indicators was a study conducted by Bollen, Mao and Zeng (2010). They attempted to analyse whether general mood on Twitter feeds could be used to predict financial indicators. They compared different SA methods to determine if the development of the mood could be used to predict the indicator. They also tested different correlation analysis methods, the Granger causality test and self-organizing, fuzzy neural networks, to determine if the closing values of the DJIA (Dow Jones Industrial Average) could be predicted. They concluded that public mood states derived from Twitter feeds can be used to predict the DJIA. “We find an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA (Dow Jones Industrial Average) and a reduction of the Mean Average Percentage Error by more than 6%” (Bollen, Mao & Zeng 2010, p. 1). They found that multiclass SA gave the most accurate results. The results in this much cited study were very encouraging and triggered a large amount of new research in the area of

PA for financial indicators using SM data. PA has also been used to predict sales volumes, elections, diseases and political and humanitarian crises.

Wei, Mao and Wang (2015) investigated the relationship between Twitter volume spikes and stock options pricing. Their study found that implied volatility increases sharply before a Twitter volume spike and decreases rapidly afterwards. Also, they found that options may be overpriced after a spike of tweet volumes. Similarly, Souza et al. (2015) tried to predict whether there is statistically-significant information between the Twitter sentiment and volume, and stock returns and volatility. They concluded that measures of the Twitter sentiment extracted from listed retail brands have a statistically-significant relationship with stock returns and volatility. Using Facebook SA, Siganos, Vagenas-Nanos and Verwijmeren (2014) observed a positive relation between sentiment on Facebook and stock market returns.

Ostrowski (2011) performed a study using relational modelling and a social dimension to predict customer behaviour. The aim of the study is to support more focused customer relationships. His findings showed that among his three chosen supervised classification methods the edge-based k-means outperformed the modular matrix approach. K-means is a clustering technique used to categorize the tuples of a dataset in different groups based on the similarities (Kumar et al. 2012, p. 28). The modular matrix approach organizes data into matrices based on user defined classes. The dimension of the matrix corresponds to the number of classes. The study compared the performance of the classification methods under various scenarios but did not provide an actual application for the customer relationship management.

Wong, Sen and Chiang (2012) have analysed whether tweets can be used to predict box office sales for movies. They accessed tweets using the Twitter API and used IMDb and

Rotten Tomatoes (RT) ratings as control data. They concluded that Twitter users are generally more positive about movies than IMDb and RT users and that Twitter reviews did not necessarily translate into predictable box-office sales. This suggests that the results can differ depending on which SN is being analysed. Our proposed study will take this into account when it comes to determining the reliability of the results.

However, other studies have challenged the results and the validity of PA. Gayo-Avello (2012) challenged the capability of Twitter analysis to predict elections. He stated that SA on Twitter data could not unambiguously distinguish between political and non-political tweets. He also found that tweets can refer to different persons, even if they make use of the same name. He cited the example of Felipe Calderón, who was a candidate for the elections for the Mexican president but also a Spanish candidate, Ramón Calderón, who was running for president of Real Madrid.

As can be seen from this, many factors influence the accuracy of predictions and can, in particular, be significantly influenced by the data collection and feature extraction steps of the process.

Method	Source	Synopsis
Data mining	Trif (2011); Kumar et al. (2012); Lim, Chen & Chen (2013); Bulysheva & Bulyshev (2012); Zeng, Li & Duan (2012)	Utilizing appropriate data mining algorithms is crucial to obtaining meaningful results. Different algorithms perform differently depending on the data types, the volume and the velocity. Back-propagation, evolutionary and statistical algorithms have been analysed. They are used for classification, aggregation, segmentation, text mining and sentiment

		analysis as well as associations.
Predictive analysis	Bollen, Mao & Zeng (2010); Tumasjan et al. (2011); Arias, Arratia & Xuriguera (2014); Asur & Huberman (2010); Wong, Sen & Chiang (2012); Wei, W, Mao, Y & Wang, B (2015); Souza et al. (2015)	A number of studies tried to use SM data for predictive analysis. Predicting box office sales on the opening weekend of a new movie, the development of the Dow Jones index, election outcomes, TV ratings and influenza rates using sentiment analysis. The studies concluded that using the right methods accurate predictions could be made. They also concluded that text mining of SM data is still very challenging.

Table 2-3: Big Data literature overview

2.2 A framework for analysing Twitter data

The literature review has shown that SN are being analysed from many different perspectives. There is a large body of literature on profit maximisation through online marketing campaigns and their return on investment. There has also been increasingly innovative research on generating new business by analysing SM data. Many studies analyse the online reputation of companies or products and performed opinion and SA. The literature review also indicated that an increasing number of companies perform SM analysis, however they often struggle due to the lack of profitable business models for their organisations.

The goal of this research proposal is to develop a framework to analyse Twitter data using Big Data analysis and to determine the economic value especially for share prices.

The study complements the current literature on SMM and Big Data theories by developing a framework for tweets and evaluating different methods. It will determine which textual binary sentiment classifier produces the more accurate results for finding correlations with share price.

While there have been many studies using SM analysis to predict share price, to the best of our knowledge no study so far has correlated opinions on companies and products to the share price of the company.

2.3 Research question and issues

Based on the literature review, 2 Literature review, conducted, the following research question has emerged:

Research question: Is there a correlation between share price and public opinion on products based on Big Data analysis of tweets?

Essentially, I argue that by using predictive analysis of Twitter tweets a monetary value can be deduced. The research question is based on the following hypotheses:

Hypothesis 1: Sentiment polarity of Twitter data can be correlated to the share price history

If a correlation can be determined as assumed in hypothesis 1, this leads to the second hypothesis:

Hypothesis 2: A correlation between sentiment polarity and share price can be used to predict share price trends

2.4 Summary

This chapter reviewed the relevant literature and identified the gaps this thesis aims to fill. Previous studies have tried to predict financial indicators. Wei, Mao & Wang (2015) used the Black–Scholes mode and Twitter spikes to predict stock option pricing. They concluded that implied volatility increases sharply before a Twitter volume spike and decreases quickly afterwards (Wei, Mao & Wang 2015, p. 271). They did not use opinion mining but used only tweet volumes. Ishijima, Kazumi & Maeda (2015) used word frequencies in news articles for opinion mining to predict the Japanese stock market. However, their methods using daily word frequencies did not perform as well as more sophisticated opinion mining techniques. Shulong et al. (2014) analysed public sentiment variations on Twitter. They used the Latent Dirichlet Allocation based clustering methods for opinion mining allowing them to detect reasons behind sentiment variations (Shulong et al. 2014, p. 1168). Nanli et al. (2014) used collective mood states on Sina, a Chinese version of Twitter, to predict the Chinese stock market. While they did not describe the details of their analysis they did use dictionary based methods for opinion mining. Hong Keel, Dennis & Yuan (2014) analysed tweets of S&P 500 companies for positive and negative opinions to predict daily stock market returns of these firms. Like Nanli they employed a dictionary based approach for sentiment analysis. Zheng & Xiaoqing (2013) used a Social Behaviour Graph based on human's online behavior for sentiment analysis on a Chinese stock forum to predict trading volumes. Porshnev, Redkin & Shevchenko

(2013) used a lexicon based approach for sentiment analysis of tweets and SVM and neural networks for predictive analysis. Evangelopoulos, Magro & Sidorova (2012) applied Latent Semantic Analysis to extract the semantic and conceptual content from tweets to make predictions on stock prices. Bollen, Mao & Zeng (2010) used a tool for opinion mining, OpinionFinder, to analyse tweets to predict the DJIA. The literature review indicated that no study applied regression based models for opinion mining. Regression models have been used for predictive analysis (Souza et al. 2015; Evangelopoulos, Magro & Sidorova 2012), but not study used LR for creating opinion time series. LR finds a single, linear decision boundary and has the advantages that:

- It does not expect the independent variables to be normally distributed
- It does not expect a linear relationship between the independent and the dependent variable
- It has a low variance
- It is less prone to over-fitting from other learning schemes such as neural networks
- It is a fast algorithm

Applying LR for opinion mining has the potential to improve classification accuracies when applied to text due to the heterogeneity and noisiness of posts on SM.

The next chapter describes the research methodology that has been adopted.

3 Research methodology

3.1 Introduction

The previous chapter discussed the ever growing literature in the area of SMM and Big Data and identified the gaps this study intends to fill. This chapter will discuss the research design and methodology, outlined in the introductory chapter in detail and justify the approach.

DM is a very iterative process and for certain tasks several methods were used. The best performing method was then chosen for the specific task. For instance, the original approach of using word frequency analysis for opinion mining did not provide satisfactory results, so a more sophisticated method based on n-gram and subjectivity analysis was chosen.

3.1.1 Data conditioning phase

3.1.1.1 Data collection of Twitter data

Usually data analysis starts with getting familiar with the domain. This is important because different areas often use different language and its own jargon. This influences how data is searched for, viz. the search terms, and how it is being mined and interpreted.

Once the domain is understood, the sources have to be defined. This step has gained in importance since the Internet has opened a whole wealth of new data sources. For this study Twitter has been chosen for several reasons:

- Twitter has a public search API to access historic posts and a streaming API to access real-time data. In contrast, Facebook has a query API limited to the 100 latest posts of a user's timeline and access to the streaming API, the "Public Feed API", is limited to a small invited research group.
- By definition all tweets are public as a result of which they can be read by anybody, unlike other SM sites, e. g. Facebook, where posts can be accessible only by Facebook friends if not made public by the user.
- Twitter tweets are limited to 140 characters. This limit does not allow for lengthy reasoning and usually posts are direct and to the point. This makes opinion mining easier since usually explicit statements are used and no complex linguistic constructs need to be processed.
- Twitter offers a powerful query API where search terms can be refined to include and exclude tweets containing certain search terms. This allows result lists to be filtered down to include only relevant tweets and the use of pre-processing techniques such as relevance filtering when constructing the queries.
- Twitter has a large, active user base. To obtain good results for PA and ML, a large volume of labelled training data is needed.

With a user base of more than 650 million Twitter accounts and an average of close to 7'000 tweets sent every second (Twitter Blogs 2015), Twitter produces large volumes of data. For this study, a popular consumer product company, Apple Inc., was chosen, since lesser known companies are less tweeted about or hardly at all.

Data was collected during the periods when there were no announcements from Apple or any conferences held where Apple was presenting in order to avoid having outliers. After any company announcement, an unusually high number of users were likely to tweet in

response and retweet in response, thereby producing an unusually high number of tweets. Outliers are instances that are considerably different from other instances in the dataset (Zafarani, Abbasi & Liu 2014, p. 141). Outliers usually falsify the result and are thus often removed. While outliers might be reflected in a financial index too, here the assumption was that a correlation is more likely to be detected using a normalized time series curve. Outliers can be handled using Big Data methods. However, this study aims to find correlations using Twitter tweets only. There are many external influencers that might influence share price such as marketing campaigns or exhibitions. Including external influencers could be a subject for future research.

When collecting data, we can either use APIs provided by SM sites for data collection or scrape the information from those sites (Zafarani, Abbasi & Liu 2014). Using the API will require writing a program or script to collect the tweets automatically. Using a screen scraper means using a program, typically a browser plugin, to simulate a user and scrape the data off the SM site. No programming is needed. Some SM sites offer different types of APIs. For instance, Twitter has a search API for historic data and a streaming API for real time data. Twitter offers a “firehose” API for 100%, a “gardenhose” API for 10% and a “spritzer” API for 1% of its real time data. If no API is provided, screen scrapers can be used to access SM data. It accesses data directly through the browser similarly like a user does, for instance a Facebook time line or a Twitter search result. Web screen scrapers are usually browser plug-ins such as NCapture (NVivo 11 for Windows 2016) or Mozenda (Mozenda Blog, 2017). There are also online tools that can be used to access the data such as Spredfast (Spredfast 2017) or Topsy (Topsy.com 2015).

For this research the NCapture screen scraper that comes with NVivo was used and the Twitter4j (Twitter4j 2017) library for collecting historic tweets. Twitter4j is a Java library to facilitate access to Twitter APIs through a Java program. It gives access to the timeline, to the timeline of friends, the search API, streaming (real-time) data, number of followers, retweets and so forth and can execute queries.

Twitter requires its users to register and create an access token to access its data programmatically. It is based on OAuth, an open standard for authorization on the Web.

Through the API, the number of retweets, the number of mentions in other tweets, the number of followers, and the time stamp and location trends can be accessed for each tweet. The Twitter API also gives access to queries.

The output of a query lists the user and the content of the post. A sample query output is shown in Figure 3-1: Twitter data query output

```
@Lisa_Mainiac_ - Samsung Galaxy Tab 3 ? ?
http://t.co/cbFk4O7o2s http://t.co/zKPh7vF08I
@electronicguid - Must See: Samsung Galaxy Note 2 vs. Galaxy Mega 6.3 Comparison
Part 1 http://t.co/PcNeywtwqL
@ForVi_ - Name: @Forvi_ The New generation has Come I AM #ReadyToMoveOn with
@Samsung_ID Galaxy S III Mini Go ! 32 ##ReadyToMoveOn
@Marmaladedays - @O2 and get a great case for it here :) https://t.co/fq8yt9yUjs
@tsriram - Samsung Galaxy Note 3 Hammer & Knife Test | Freak.
http://t.co/bAZtU6DFZa
@GoulartIzabel - Micro USB MHL TO HDMI 1080P HD TV Cable Adapter For Samsung Galaxy
S3 S4 Note 2: Price 0.76 USD (9 Bids) End... http://t.co/n5VAHZpbVW
@Benjovi23 - If you have a Samsung galaxy device don't send texts with the smiley
faces its changes it to picture messaging and you will get charged
http://t.co/iwszKpQtiy 15.12 10:36
…
```

Figure 3-1: Twitter data query output

The Twitter search API is not idempotent, meaning, it will not always return the same results. Two consecutive, identical Twitter searches, one through the API and one using the Web search may render different results. The documentation of the Twitter Search

API documentation states that the API renders data based on relevance (Twitter Developers 2015). This is only a problem if the feature distribution in different query result sets varies significantly. Since according to the Twitter documentation the tweets are selected randomly, the distribution should be more or less even.

In this study, the tweets were stored in an Amazon S3 bucket, a storage service offered by the Amazon cloud. The Twitter search API has the possibility to refine a search by excluding words, including or excluding retweets, or giving a start and end date. A smiley can be added to retrieve only tweets with a positive or negative attitude. For instance, the following search retrieves all tweets containing the words “iPhone S6” since 1st July 2015, excluding tweets containing http or https as possible spam (Bollen, Mao & Zeng 2010, p. 2), excluding retweets and with a positive attitude:

```
iPhone S6 since:2015-07-01 -http -https +exclude:retweets :)
```

Figure 3-2: Twitter search

This has the advantage that part of the data pre-processing can already be done during the data collection phase. For instance, excluding retweets in the query can handle data deduplication during data collection, at least partly as has been found in this study. The query can be called programmatically through the API or through the advanced search in Twitters web site. The search results can then be collected using NCapture. In this study, both methods were used. It has to be noted that searching through the API and through the Web interface showed to return different result sets. Twitter returns a sample of random tweets for a given time period and query according to the Twitter documentation (Twitter Developers 2015). No additional information about how Twitter actually selects the random tweets was documented.

It should be noted that according to Big Data principles data cleansing such as spam removal should not be necessary. However, for the algorithms used in this study, a purified data set had a better predictive performance from a dataset that did not go through the cleaning steps. The data cleansing stage can have a direct effect on both the accuracy and speed of the entire resolution process (Ting, Wu & Ho 2010, p. 75). Also, better features will result in better trained learners. Since ML map input(s) *x* to output(s) *y*, the closer the input vector *x* is to *y*, the easier the task for a ML scheme becomes. The closer the input and output vectors are, the faster the learner converges during training.

3.1.1.1.1 Datasets

For training, several datasets were tested. For instance a data set containing opinions about Apple, Google and Microsoft was used but did not show satisfactory classification accuracies when the trained classifiers were evaluated. These datasets were not used any further in the testing.

For this study, only tweets about Apple, the iPhone and the iPad were used. The iPhone and iPad are contributing most to Apples revenue Pramuk 2015). Apple has a number of other products such as the Apple watch, the iMac, the MacBook and many others. An obvious choice would be to collect all tweets about Apple and all Apple products. However, Apple also sells accessories to their products, lesser known products such as the Apple Pencil or the Xsan Storage Area Network and services. However, only the products with the biggest contributions to revenue were analysed. For instance, the Apple Pencil only contributes marginally to revenue. Also, considering service such as the “Apple Volume Purchase Program for Business”, a search returned only one result. Apple also has contracts with some large corporations where the details are not public and no information about revenue is publicly available. For this reason, this study

focuses on tweets directly about Apple, the iPhone and the iPad. Many possible combinations could have also be considered, for example all Apple, iPhone and iPad tweets together. However, testing all combinations for correlations would exceed the scope of this study and should be subject for future research.

3.1.1.2 Data pre-processing

During the data pre-processing stage, the data is passed through a relevance filter. Irrelevant data such as stop words, punctuations such as brackets or semi colons and smileys are typically removed. ML techniques such as Logistic Regression or the Naïve Bayes classifier used in this study use a bag-of-words approach and do not require stop words to be removed. There is no definitive list of stop words. For example, “isn’t” is sometimes removed as stop word, but in opinion mining it is a mood polarity shifter that can change the entire meaning of the statement. To avoid spam, text that matches the regular expression “http” and “www” was removed.

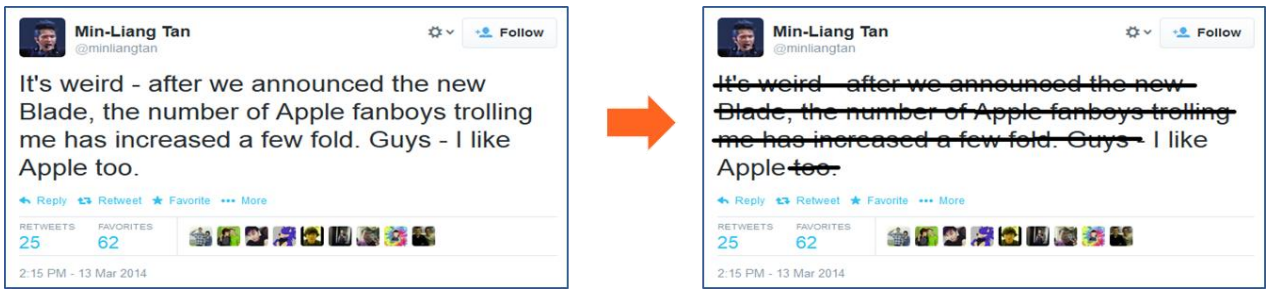


Figure 3-3: Relevance filtering of tweet

To get clear statements, a data set containing only tweets with subjective statements was created, viz. data records that contain opinions, not fact statements. Tweets with fact statements were discarded in this set. For instance, “I like the new Tesla” is a

subjective statement expressing a sentiment, “The new iPhone comes in black and silver” is a fact statement and has to be removed.

Data deduplication is an important data pre-processing step. Twitter searches can have many duplicates even though the “-filter:retweets” parameter was added to every query. The similarity between the documents is usually computed by one of the distance or similarity measures such as cosine similarity, Euclidean distance, Pearson correlation coefficient, Jaccard index, Kullback-Leibler divergence and many others (Zlacky et al. 2014, p. 161). To remove duplicates, in this study the Jaccard index was used. The Jaccard index operates at a token level and compares two strings by first tokenizing them and then dividing the number of common tokens by the total number of tokens (Baldwin & Dayanidhi 2014, p. 427). The Jaccard index is a very efficient way of comparing strings and has been used in many studies (Zhong et al. 2016; Zafarani, Abbasi & Liu 2014; Vesdapunt, Bellare & Dalvi 2014; Hangya & Farkas 2013). The deduplication was run with a proximity cut off of 0.5 where a proximity of 1 is a perfect match. All rows above proximity were eliminated. The threshold was determined empirically by looking at results with thresholds from 0.3 - 0.7. Deduplication is needed to reduce redundancy and improve the performance of the algorithm. If every row has to be compared to all other rows, the algorithm is *O(n^2)*. That is why it is not suited for large datasets.

To pre-process the tweets, for example for data deduplication and subjectivity analysis, the Natural Language Processing (NLP) toolkit LingPipe (LingPipe 2015) was used. LingPipe has implemented all common NLP techniques such as part-of-speech tagging, named entry recognition, word sense disambiguation, expectation maximisation and many more. It is open source and implemented in Java.

After filtering out all the relevant information, the tweets are ready for SA. The pre-processing tasks depend highly on the analysis problem and can vary depending on the type of data, the problem at hand and the results that the analysis is supposed to reveal. For this study several datasets were analysed to compare their performance with different ML schemes.

3.1.2 Predictive analysis phase

3.1.2.1 Data classification

After the data has been pre-processed, it is ready to be classified. Classification can be binary, usually positive or negative, or multiclass, for instance a Likert scale. In this study the data was classified into positive and negative tweets over a time frame of two months. The binary sentiment polarity was then represented as a time series. Modelling for time series is conceptually similar to other modelling problems, but one major distinction is that usually the next value of the series is highly related to the most recent values, with a time-decaying importance in this relationship to previous values (Wu & Coggeshall 2012, p. 173).

The tweets were classified using supervised ML techniques. We interpret ML as the acquisition of structural descriptions from examples (Witten, Frank & Hall 2011, p. XXI). It allows a system being trained using examples and adding human judgement to correct the incorrectly classified tweets. To train a learner, after each learning cycle the error is calculated using a loss function such as the negative log-likelihood or mean squared error. During training the loss function is minimized until a minimum is reached. ML, then, is about making computers modify or adapt their actions (whether these actions are

making predictions, or controlling a robot) so that these actions get more accurate, where accuracy is measured by how well the chosen actions reflect the correct ones (Marsland 2009, p. 5). ML can be useful in situations in which producing rules manually is too labour intensive (Witten, Frank & Hall 2011, p. 25). ML techniques are also adopted when a problem cannot be adequately solved using a simple (deterministic), rule-based solution or when it does not scale. Spam filtering is probably one of the most prevalent applications of ML. It cannot be done manually due to the large of volumes of spam and it does not scale since constantly new forms of spam appear and programming new rules cannot keep up with the speed at which new forms of spam surface. Spam filtering is often done using techniques based on the Bayesian theorem. For each email the probability that it is legitimate or spam is calculated. If there is a high probability that the email is unsolicited, it is filtered out by the spam filter.

To analyse and classify the tweets, several ML tool kits were evaluated. For this study, as for NLP, LingPipe was used. LingPipe is an open source NLP library that has implementations of ML techniques that can analyse a body of data. It has implementations of classification and clustering algorithms. It is well documented, has been used in many real-world problems (Carpenter 2007; Konchady 2008; Denecke 2008) and was easy to use. The frameworks that were evaluated were LingPipe (LingPipe 2015), Mallet (MALLET 2017), WEKA (Weka 3 2015), RapidMiner (RapidMiner 2017), NLTK (Natural Language Toolkit 2017) and Amazon ML (Amazon Machine Learning 2017). The evaluation of the tools focussed on popular open source frameworks except for Amazon ML. MALLET is an open source ML framework that has implementations of many popular ML algorithms. However, the documentation showed to be very incomplete and the online community seemed to be smaller from the LingPipe

community. WEKA is a ML framework written in Java, but contrary to MALLET has a graphical user interface. WEKA has no NLP capabilities. RapidMiner is a commercial DM and ML tool that has a free, open source version. The open source version has a limit of 10’000 data rows (RapidMiner 2017) which is too small for this study. NLTK is a very popular NLP toolkit written in Python. Its functionality is comparable with the functionality of LingPipe. It has NLP and ML capabilities. Amazon ML was the only non-open source toolkit evaluated. At the time of writing there was no documentation of which ML algorithms Amazon ML uses. LingPipe was selected over the other tools because it is well documented. Since the data was collected using Twitter4j and Java, using a Java library was a natural choice over a Python based library. As stated before, the Documentation of MALLET was very incomplete which made its use very difficult. WEKA has no NLP capabilities; RapidMiner in its free edition was too limited and Amazon ML did not document any of its used algorithms.

Generally speaking, ML schemes are well documented in literature and many good implementations, open source and closed source can be found. The focus of the evaluation was on ease of use, availability of documentation and on how easy it could be integrated into the framework used for this study. That is the main reason why LingPipe was selected over NLTK, which is written in Python, not Java.

Figure 3-4: Machine learning cycle shows a ML learning cycle using LingPipe.

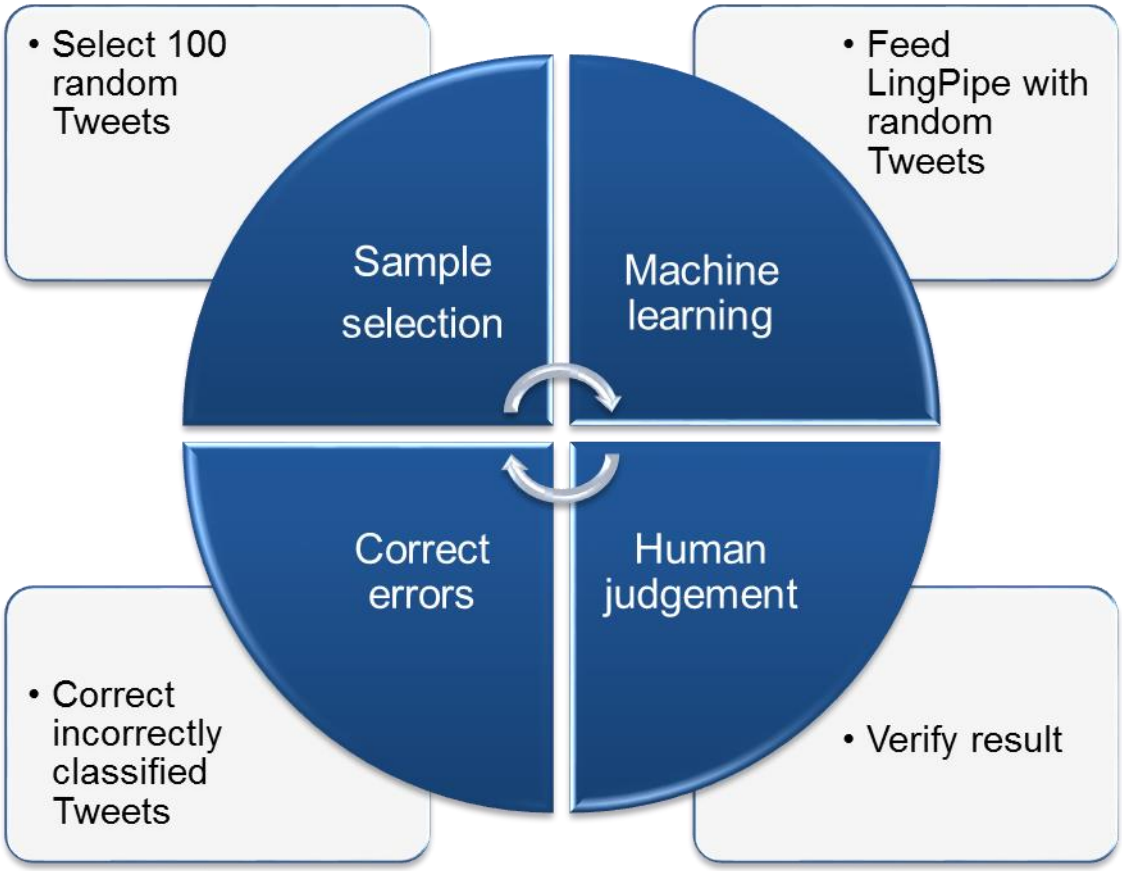


Figure 3-4: Machine learning cycle

Ultimately we want to find a decision function *f*, that classifies tweets into labels $X=\{x_1, x_2,\ldots x_n\}$, such that $f:X \rightarrow \{P,N\}$, predicts if a tweet is positive *P* or not *N*. This is a binary classification problem since we have two class labels, *f* is called classifier. It should be noted that many classifiers output the probability *Pr*, that a tweet x_i with corresponding label y_i belongs to class *j* (Wlodarczak, Soar & Ally 2015, p. 380):

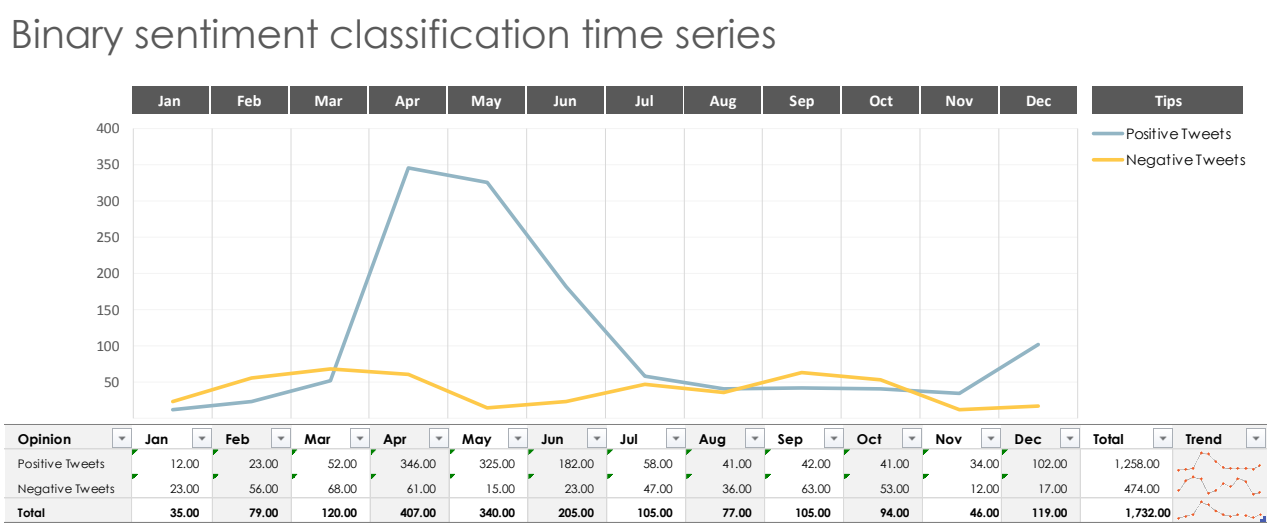
$$\Pr(x_i | y_i = j)$$

Equation 3-1: Classification probability

In this study a method called cross-validation was used. In cross-validation, first the labelled data is separated into a set of training data and a set of testing data. The data is divided randomly into *n* folds, where *n* is usually somewhere between 5 and 10. The model is trained with *n*-1 folds, then, the model is tested using the holdout fold. This method is called *n*-fold cross-validation. There are usually several iterations until the result converges. Converging means the result is not changing anymore. The result is evaluated using different measures. A popular measure is the classification accuracy, the ratio between the correctly classified and the total number of tweets. Other measures used were the F score and Kappa statistics.

There are many measures for the performance of a binary classifier. The Receiver Operating Characteristic (ROC) is a graphic plot that illustrates the varying discrimination threshold. Precision–Recall (PR) measures the relevance of a classification. The performance measures will be explained in more detail in the research methodology chapter.

This process of training and testing is repeated for all the classification algorithms to determine which one has the best classification performance. Experience shows that no single ML scheme is appropriate to all DM problems (Witten, Frank & Hall 2011, p. 403). LingPipe has several evaluator classes that can be used to compare the performance of the different algorithms to select the most effective one. After evaluation, a set of new, unseen data was run against the trained algorithm to create a binary sentiment classification time series of two month worth of tweets.



Opinion	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total	Trend
Positive Tweets	12.00	23.00	52.00	346.00	325.00	182.00	58.00	41.00	42.00	41.00	34.00	102.00	1,258.00	
Negative Tweets	23.00	56.00	68.00	61.00	15.00	23.00	47.00	36.00	63.00	53.00	12.00	17.00	474.00	
Total	**35.00**	**79.00**	**120.00**	**407.00**	**340.00**	**205.00**	**105.00**	**77.00**	**105.00**	**94.00**	**46.00**	**119.00**	**1,732.00**	

Figure 3-5: Binary sentiment classification time series

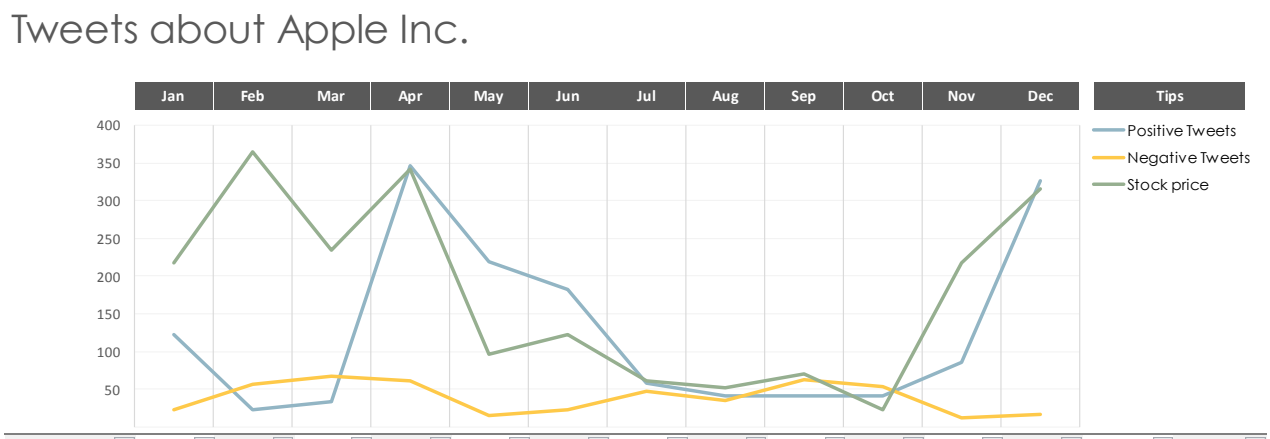
3.1.2.2 Correlations

To determine if predictions can be made, correlations have to be detected and validated; in this case is tweet there a correlation between share price and public opinion on Twitter? In other words, does the Twitter mood time series have statistically relevant predictive information about the financial time series?

A popular method to analyse the temporal relationship of the behaviour is the Granger causality test. It evaluates the correlation between two lagged time series. However, even Granger causality is not causality in a deep sense of the word because it is also based only on numeric predictions (de Siqueira Santos et al. 2013, p. 11). The traditional linear Granger test has been widely used to examine the linear causality among several time series in bivariate settings as well as multivariate settings (Bai, Wong & Zhang 2010). The original Granger tests examined the linear causality among several time series in a bivariate and multivariate setting (Wlodarczak, Soar & Ally 2015, p. 4). However, many real world applications are nonlinear and variances have been

developed (Hiemstra & Jones 1994). For this study, the R implementation in the lmtest package was used (Package lmtest 2017). The lmtest package only supports Granger in a bivariate setting, however for this study no multivariate data was analysed.

If a correlation has been detected, the latency has to be determined. The correlation latency is the delay between the change of the sentiment polarity and the change in share price. The correlation analysis was thus performed using different time lags. To determine if correlations exist between Twitter opinions and share price, the sentiment time series was compared against the stock price chart of the analysed company. Figure 3-6: Tweets about Apple Inc. and AAPL share price shows the Twitter mood time series and the Apple Inc. share price.



Opinion	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total	Trend
Positive Tweets	123.00	23.00	34.00	346.00	219.00	182.00	58.00	41.00	42.00	41.00	86.00	327.00	1'522.00	
Negative Tweets	23.00	56.00	68.00	61.00	15.00	23.00	47.00	36.00	63.00	53.00	12.00	17.00	474.00	
Stock price	218.00	365.00	234.00	341.00	97.00	123.00	61.00	52.00	71.00	23.00	218.00	315.00	2'118.00	
Total	**364.00**	**444.00**	**336.00**	**748.00**	**331.00**	**328.00**	**166.00**	**129.00**	**176.00**	**117.00**	**316.00**	**659.00**	**4'114.00**	

Figure 3-6: Tweets about Apple Inc. and AAPL share price

If a correlation was detected visually, the detection had to be automated. A correlation exists, if a pattern in one time series repeats in the other time series.

To find correlations, the R statistics framework was used and the lmtest package and it's the Granger causality test implementation. The test is simply a Wald test comparing the unrestricted model—in which *y* is explained by the lags, the order, of *y* and *x*—and the restricted model—in which *y* is only explained by the lags of *y* (Package lmtest 2017). The inputs for the causality tests were:

1. A time series of two month of stock market values, for example Apple stock price AAPL
2. Permutations of mood time series

The analytical framework was written in the Java programming language and used the LingPipe NLP toolkit for data pre-processing and analysis. The R statistics environment and programming language was used for correlation analysis. An R script was developed to run the Granger causality tests.

3.2 Data analysis framework

For this study, a data analysis framework was developed using Java and R. It consists of a data collection class that uses Twitter4j (Twitter4j 2017) to access the Twitter REST API. Some helper classes were developed, for instance for data transformation of tweets from text to CSV format. The data went through data deduplication and basic subjectivity pre-processing steps using LingPipe for NLP. LingPipe was also used for feature extraction. In the case of the naïve Bayes classifier, a bag-of-words was created. For LR a tokenizer from LingPipe was used to extract n-grams.

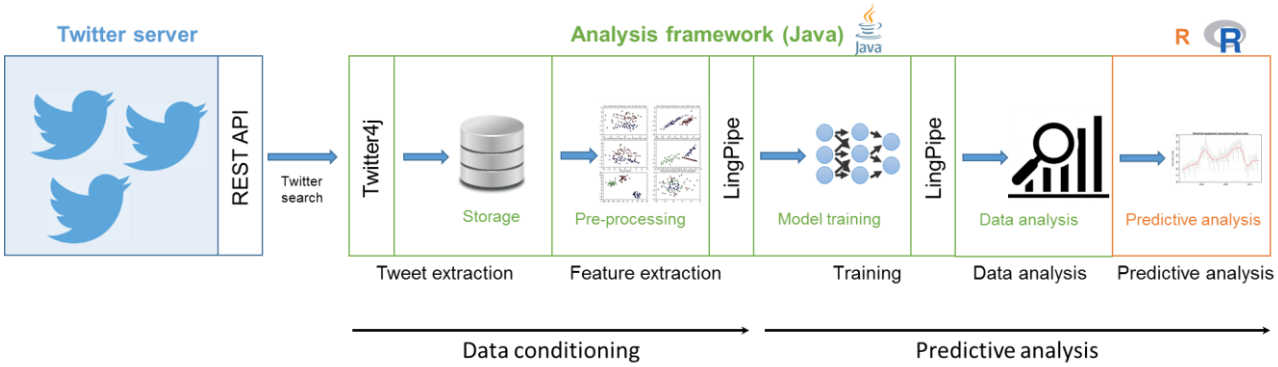


Figure 3-7: Data analysis framework

The framework is generic. Except for the data collection, it could be used to analyse posts from other SM sites. However, since tweets are limited in character size, it would have to be tested with longer posts to verify its suitability for other sites allowing posts exceeding the 140 character size.

3.3 Justification for the paradigm and methodology

ML techniques have been widely used for DM. They have shown to be effective for many DM problems (Xinyu, Youngwoon & Suk young 2015; Wei, Mao & Wang 2015; Tsakalidis et al. 2015). ML techniques are domain independent and can be used for text mining, multimedia mining, decision support, predictive maintenance or fraud detection. Many approaches have been applied for SA (Zlacky et al. 2014; Solakidis, Vavliakis & Mitkas 2014; Shulong et al. 2014). Topic models such as Latent Dirichlet Allocation (LDA) (Zlacky et al. 2014) or term weighting such as Term Frequency and Inverse Document Frequency (TF/IDF) (Gokhale et al. 2014) have been widely used. Many more methods exist. The decision to use supervised ML techniques was taken for following reasons:

- The class label was known; in unsupervised methods such as LDA observations are assumed to be caused by latent variables
- The causal relation between the input and output observations is straightforward, the causal gap is small
- There is an abundance of data, about 6’000 tweets per second on average
- ML techniques usually have good generalization performance as long as there is enough data
- ML has an emphasis on predictive models, statistics is more focussed on explanatory or inferential models
- The problem at hand is linearly separable
- Naïve Bayes and Linear Regression have been widely used and can be very effective for many data analysis problems, especially for text processing tasks

ML models fall into the category of context reasoning decision models. Context reasoning can be defined as a method of deducing new knowledge, and understanding better, based on the available context (Perera et al. 2014, p. 432). They support the decision making process by making predictions about imminent machine failure, customers who are likely to switch or possible security breaches in case of a cyber-attack. There are a large number of different context reasoning decision models, such as decision tree, naive Bayes, hidden Markov models, support vector machines, k-nearest neighbour, artificial neural networks, Dempster-Shafer, ontology-based, rule-based, fuzzy reasoning and many more (Perera et al. 2014, p. 432). The requirement for context reasoning emerges due to imperfections in raw data. When collecting bulk data such as tweets, there are likely to be off topic tweets, duplicates, outliers, missing data or

misinterpreted tweets. ML techniques mitigate this problem by giving high-level context deductions from a set of contexts.

ML is inspired by natural learning and as such it is well suited for real world problems such as opinion mining. But it should be noted that human learning is supervised and unsupervised, but dominantly unsupervised. Humans learn mostly from experience, not from labelled data.

The Granger causality test (Souza et al. 2015) was used in assessing whether there are anticipatory or lagged effects in time series. The hypothesized causal variable in time series *X* must have unique information about the dependent variable in time series *Y*. While Grangercausality has no real meaning for causality, it is well suited for analysing time precedence in explorative studies and it is widely used in economics. The causal relationship must not be simultaneous but defined with a lag, which is an assumption in this study.

3.4 Ethical considerations

The use of data—particularly data about people—for data mining has serious ethical implications, and practitioners of data mining techniques must act responsibly by making themselves aware of the ethical issues that surround their particular application (Witten, Frank & Hall 2011, p. 33). Very little is understood about the ethical implications underpinning the Big Data phenomenon (Boyd & Crawford 2012, p. 672). In this study, a lot of opinionated tweets about Apple were collected. However, the data was collected anonymously and no personal information was extracted. Nevertheless, using this framework, there are several considerations. Often reidentification of anonymised data is possible. For instance, over 85% of Americans can be identified from publicly available

records using just three pieces of information: five-digit zip code, birth date (including year), and sex (Witten, Frank & Hall 2011, p. 33). Tweets are by definition public for everyone to read. There seems to be no privacy issue. However, often people are not aware of what information they are disclosing that is implicit. For instance, one study could determine sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender only based on Facebook likes in some cases with accuracies of more than 80% (Kosinski, Stillwell & Graepel 2013, p. 5802). So Twitter users, without being aware are revealing information about their interests, their opinions and views. The information can be used for targeted advertisements or marketing campaigns. Oppressive regimes can also use the info to find and persecute dissidents or critical journalists. Users of SM are often ignorant about what information can be deduced from what they publish. Privacy-preserving DM techniques have been proposed (Vatsalan, Christen & Verykios, 2013; Boyd & Crawford 2012; Agrawal & Aggarwal 2001; Agrawal & Srikant 2000). However, due to the complexity of the subject they are often not applied. It should be noted that collected data might be used in ways that go far beyond what was originally planned when it was collected.

When applied to people, data mining is frequently used to discriminate—who gets the loan, who gets the special offer, and so on (Witten, Frank & Hall 2011, p. 33). Discrimination based on opinions in tweets is only one possible form of abuse. Positive opinions often mean profits and fames for businesses and individuals, which, unfortunately, give strong incentives for people to game the system by posting fake opinions or reviews to promote or to discredit some target products, services, organizations, individuals, and even ideas without disclosing their true intentions, or the

person or organization that they are secretly working for (Liu 2012, p. 123). Writing fake opinions is called opinion spam. Twitter not only offers a search API, but also an API to automatically create new tweets (Twitter Developers 2017). It could be used to produce bulk fake opinions. Ultimately, making predictions based on Twitter might be considered insider information.

3.5 Conclusions

This chapter elaborated the research methodology. The data collection and analysis steps were described and some ethical concerns were raised. Twitter data was collected using the Twitter API using a program written in Java, and NCapture, a screen scraper. The data was then purified using basic subjectivity analysis to find only opinionated tweets and data deduplication. Several models were trained using the different datasets; the original data set, the deduplicated set and the set that went through basic subjectivity analysis. The models were evaluated for their performance and the best performing models were used for correlation analysis. Correlation analysis was done using the Granges causality test and R. The next chapter will describe how the data in this study was collected, processed and interpreted in detail.

4 Data analysis

4.1 Introduction

The preceding chapter elaborated the research methodology. This chapter describes the data analysis methods used for this study. It builds upon the research methodology described in the previous chapter and describes in detail how the data was collected, pre-processed and analysed.

4.2 Data mining

DM comprises two phases, the data conditioning phase and the PA phase. For this study, tweets were collected and analysed. The steps and tasks used are summarized in Figure 4-1: Twitter data mining steps.

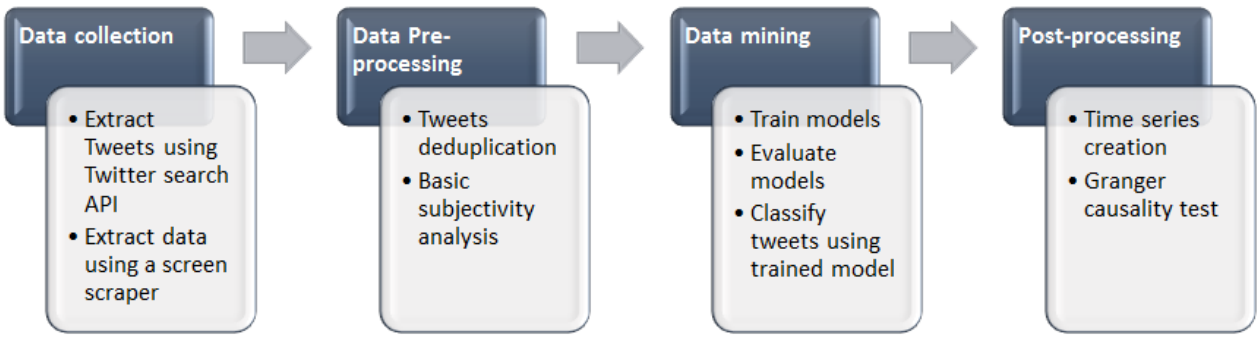


Figure 4-1: Twitter data mining steps

4.2.1 Data conditioning phase

In the data conditioning phase, tweets about Apple Inc., the iPhone and the iPad were collected using the Twitter search API and a screen scraper, NCapture from 1st of December 2015 to 31st of December. During this time period no announcements from

Apple or end user fair took place that would have boosted the number of tweets. The aim of this study is not to analyse the influence of external events on Apple share price. Selecting a period without major Apple related events avoided having outliers in the datasets. Not having outliers can be easily verified using Topsy. Figure 4-2: Number of tweets about Apple, the iPhone and the iPad shows that over the period from 2nd of November 2015 to 2nd of December 2015 no outliers in terms of number of tweets were detected.

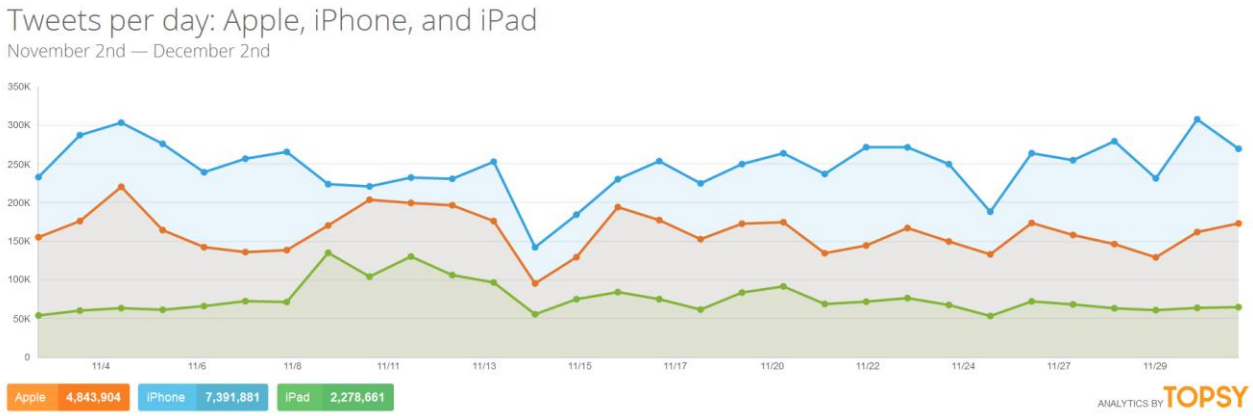


Figure 4-2: Number of tweets about Apple, the iPhone and the iPad

Apple's flagship product and main source of profit, the iPhone, is contributing most to its revenue (Pramuk 2015). It is to be expected that iPhone sales influence share price the most.

Also, a query for iPhone is unambiguous. It most likely only returns tweets specifically related to the iPhone. For instance, a query for Galaxy, Samsung's flagship smartphone suite, would most likely also return tweets about galaxies of the universe. Samsung is a conglomerate that produces other electronics such as displays, TVs, chips, hard drives etc. The Galaxy smartphone range is only one of many products that contribute to company results. This makes Apple ideal for correlating tweet trends about their products

and share price. Using for instance “Galaxy”, Samsung’s smartphone, as a search term, a query will more likely also return tweets about Galaxies in the universe. This means that additional NLP pre-processing steps such as word sense disambiguation or part-of-speech tagging are required. Generally speaking, data pre-processing is domain specific and the steps depend on the data. In contrast, ML is not domain specific and ML techniques can be used for DM in any domain.

SM companies have changed access to their sites several times in the past and it is possible that access will change again and some of the statements will become outdated. At the time of writing, Twitter restricted access to 180 queries per 15 minutes for users represented by access tokens (Twitter Developers 2015). This means that data collection has to run over a certain period of time to get the tweets. If the API quotas are changed and get more constraint, data collection will have to be run over a more extended period of time.

4.2.1.1 Access method

There are several methods to access Twitter tweets. The first is using Twitter’s search API. For this study the Java open source library Twitter4j (twitter4j 2015) was used. It allows accessing historic tweets through the query API as well as real-time data by implementing its TwitterStream interface. This has the advantage that the same query strings can be used for historic and for streaming data. Twitter4j implements a Java wrapper for Twitter’s v1.1 REST API. REST is an acronym for Representational State Transfer and is an architecture style of the World Wide Web.

The second method used was using the NVivo screen scraper plugin, NCapture. The advantage of using a screen scraper is that no API needs to be used and no

programming is required. The captured tweets can be imported into NVivo for qualitative analysis. NVivo is a computer assisted qualitative data analysis software (CAQDAS). CAQDAS has been seen as aiding the researcher in her or his search for an accurate and transparent picture of the data whilst also providing an audit of the data analysis process as a whole - something which has often been missing in accounts of qualitative research (Welch 2002, p. 1). Using NVivo has the disadvantage that only certain data formats such as XML are supported. A learning library such as LingPipe requires, for instance, Coma Separated Value (CSV) format.

The other two possibilities, using Web sites for data collection or buying data from brokers such as DataSift (DataSift 2015), were not evaluated since they are very costly. SM data can also be accessed through web tools such as Topsy (Topsy.com 2015) or Gnip (Gnip 2015). Web-based SM analysis tools often offer free, basic functionality and premium, subscription-based functionality. Web-based social data analysis tools that rely on public discussion to produce hypotheses or explanations of patterns and trends in data rarely yield high-quality results in practice (Willett, Heer & Agrawala 2012, p. 227). Topsy was used just to query the daily tweet volumes for every query. This allowed estimating for how long a query has to be run to get, for instance, one month of Twitter data. The number of daily tweets and the 15 minutes query quotas as defined in the Twitter search API can be used to get an estimate for how long a query has to run given the quota to get a data set of a day of tweets. Topsy was bought by Apple and has been discontinued now.

4.2.1.1.1 Queries

A query is a request for information retrieval. Twitter supports query operators that modify the behaviour. For instance, tweets containing certain words can be omitted. The API also supports logical operators such as the disjunction operator OR, meaning “All tweets containing *expression 1* OR *expression 2*”. The search can also restrict the time window of the tweets and exclude retweets. For instance, the query in Table 4-1: Twitter query for Apple

```
Apple -http -https -www lang:en since:2015-04-01 until:2015-04-02 +exclude:retweets
```

Table 4-1: Twitter query for Apple

retrieves all tweets containing the word “Apple”, not containing the words “http”, “https” and “www”, only tweets in the English language from April 1st 2015 to April 2nd 2015, excluding retweets. “http”, “https” and “www” are excluded since many tweets that contain URLs are spam (Bollen, Mao & Zeng 2010, p. 2).

The Twitter search API has a parameter, an emoticon, ":)" to return only tweets with a positive, or ":(" to return only tweets with a negative attitude. However, it merely checks for the presence of a positive or negative smiley and is therefore not a real sentiment classification method. For instance, a tweet *"I miss my iPhone :-("* has a negative attitude but should be interpreted as a positive sentiment towards the iPhone.

The queries used in this study are listed in the table in Appendix 6.2 Queries.

4.2.1.2 Data collection

Twitter Data was collected using the Twitter search API and NVivo's NCapture screen grabber. The NVivo qualitative data analysis tool was used to collect data and do an initial analysis. NVivo has a screen grabber plugin, NCapture, for Google Chrome and MS Internet Explorer. NCapture was used to capture tweets through the Twitter web search and imported into NVivo for data pre-processing. The volumes captured through NCapture and through the API were about the same.

The NVivo 11 Plus edition also has quantitative analysis capabilities. For instance, the NVivo Plus version can do word frequency analysis. It also has cluster, SA and topic classification capabilities among others. NVivo does SA only at the sentence, paragraph or cell level. No information on the methods used for SA is documented. It seems that NVivo does SA based on word frequencies. There are no performance statistics and a verification using human judgement showed that the results are not satisfactory. A corpus of about 167,400,000 tweets was collected and classified using NVivo. NVivo has an auto code functionality "Identify sentiment" that does multiclass sentiment classification. It classifies tweets into four categories:

Very negative, moderately negative, moderately positive and very positive

However manually checking the results showed many false positives and false negatives. For instance, *"Dads new job gave him an iPhone 6s and I'm super jealous"* was classified as very negative. Also, the SA is done on the entity level, not the aspect level. For instance, *"An iPhone without the case is so dope"* was classified as very negative. The sentiment classifier seems to only look for positive or negative sentiment words and often misclassifies the whole sentiment polarity. NVivo, according to the

documentation, performs SA based on single sentiment words. "It is important to understand that this tool does not classify content according to sentiment. It does not take each piece of content and rate it on a Likert sentiment scale. It looks at the sentiment of words in isolation—the context is not taken into account." (NVivo 11 for Windows Help 2015).

However, NVivo was useful for pre-evaluating the data. Analysing the collected data in NVivo revealed that "+exclude:retweets" did not remove all retweets and data deduplication was still necessary. Also, only few tweets really use hashtags to highlight the subject of the tweets. This explained why queries using the hashtag, e. g. “#iPhone”, did not improve the results of the predictive model. Figure 4-3: Tweets imported in NVivo shows an import of tweets captured by NCapture in NVivo:

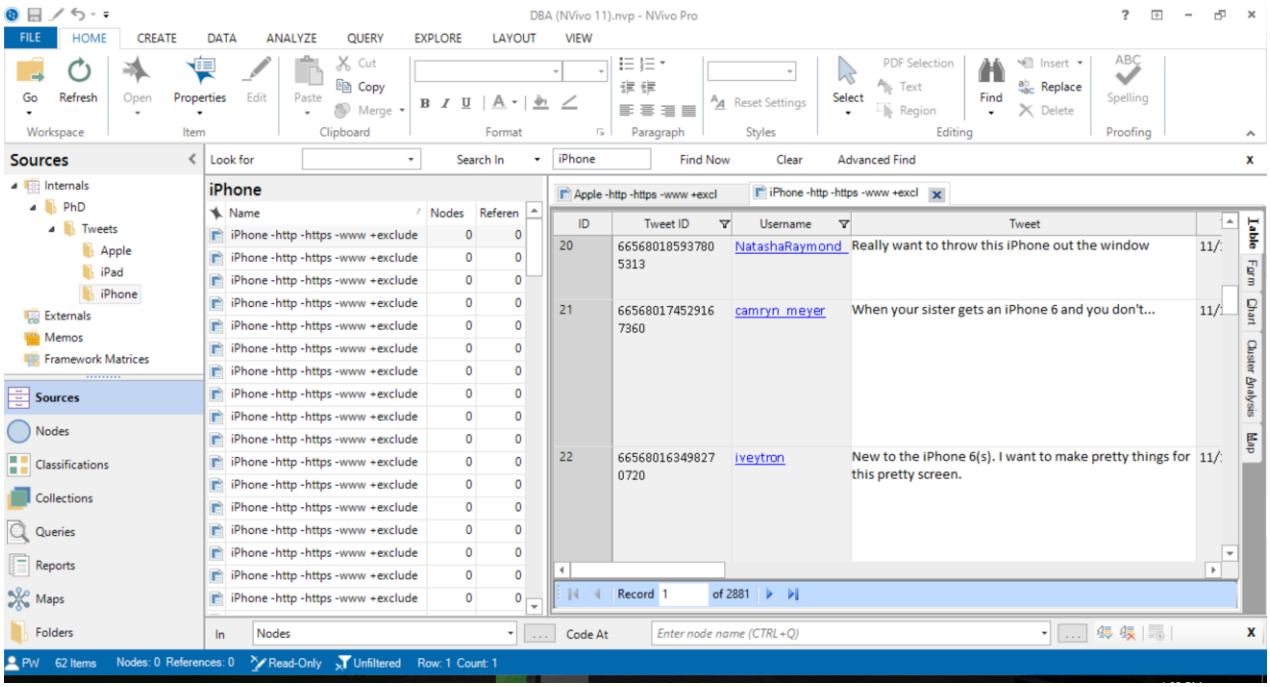


Figure 4-3: Tweets imported in NVivo

At the time of the study, Twitter limited access to its Search API to 180 queries every 15 minutes and 1600 results per query. Also, the API only returns tweets no older than 10 days. So the data collection was performed over a period of two months. A corpus of more than a quarter of a million (253,237) tweets (Apple: 85,516, iPad: 82,025, iPhone: 85,696) for the period 1st November 2015 to 31th December 2015 was collected. This body was subjected to data pre-processing and after pre-processing for training of the learning algorithms.

It's important to note that the Search API is focused on relevance and not completeness. This means that some tweets and users may be missing from search results (Twitter Developers 2015). The Twitter Search API is part of Twitter's v1.1 REST API. It allows queries against the indices of recent or popular tweets and behaves similarly to, but not exactly like, the Search feature available in Twitter mobile or web clients, such as Twitter.com search (Twitter Developers 2015). The Search API is not a complete index of all tweets, but instead an index of recent tweets. At the moment that index includes between 6-9 days of tweets (Twitter Developers 2015).

For this reason, two different methods of accessing historic tweets were used, one based on the REST API and one using the NVivo screen scraper.

The Apple share quotes were collected from the Nasdaq web site (Nasdaq 2016). Figure 4-4: AAPL quotes over a period of 2 month shows the AAPL quotes over a period of 2 months, November 2016 to December 2016 from the Nasdaq web site:

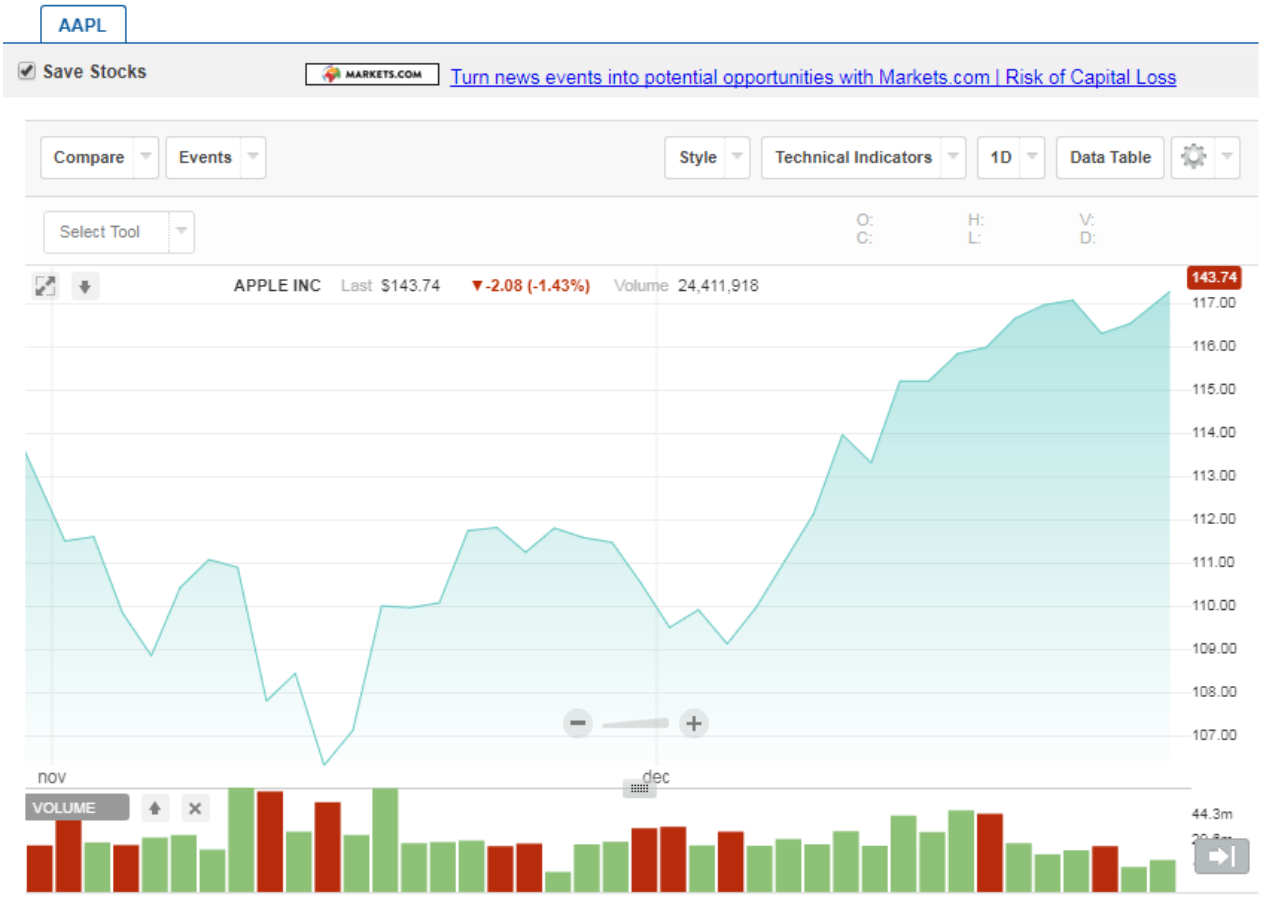


Figure 4-4: AAPL quotes over a period of 2 month

The quotes can be exported to an Excel spread sheet which allows for easier processing. *Table 4-2: Excel export of AAPL quotes* shows an excerpt from an Excel export:

Date	Open	High	Low	Close	Volume
2015-07-02	126.43	126.69	125.77	126.44	27171180
2015-07-06	124.94	126.23	124.85	126	27972950
2015-07-07	125.89	126.15	123.77	125.69	46737090
2015-07-08	124.48	124.64	122.54	122.57	60609830
2015-07-09	123.85	124.06	119.22	120.07	78291510
2015-07-10	121.94	123.85	121.21	123.28	61292800

Table 4-2: Excel export of AAPL quotes

4.2.1.3 Data pre-processing

The data pre-processing tasks include data deduplication and basic subjectivity analysis to remove tweets without opinions. Data cleaning refers to the pre-processing of data to remove or reduce noise (by applying smoothing techniques, for example) and the treatment of missing values (e.g., by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics) (Han, Kamber & Pei 2011, p. 289). Many classification algorithms have mechanisms for handling noisy or missing data. However, cleaning data before analysis can improve the learning process. Irrelevant data such as duplicates or data that does not contribute to the result have to be eliminated, for example filtered for relevance. Data transformations, such as converting the data from one format to the other, were also performed but this will not be described in detail as this only supports the analysis process but has no effect on the result. For instance, the collected tweets were transformed from plain text into Comma Separated Value (CSV) files, since this format is supported by LingPipe. To transform the data, some helper classes were developed as part of the data analysis framework.

There are two types of sentences, “objective” and “subjective” sentences. For instance, *“The new iPhone comes in two colours, black and marine blue”* is an example of an objective statement whereas *“The new iPhone is great”* is a subjective statement since it expresses an opinion, and not a fact. SA aims to extract subjective information, opinions, and thus only tweets containing subjective information need to be analysed, and objective tweets have to be filtered out. Subjectivity/objectivity identification can be a difficult task since texts can have both subjective and objective information. Also subjectivity/objectivity classification largely depends on the definition of subjectivity. For

instance, *"The new iPhone is expensive"* can be regarded as an objective statement or as a negative opinion. On SM, microblogs contain mostly subjective information of around 82.9% (Petz et al. 2014). Nevertheless, opinion mining performs better when objective sentences are removed (Pang & Lee 2004). Since tweets are short, the whole tweet was removed in this study if it contained objective sentences to create a data set with only subjective statements. To remove objective tweets, basic subjectivity analysis was performed.

Data deduplication was performed using the Jaccard index, also called Jaccard distance or Jaccard similarity coefficient. Data deduplication is a Natural Language Processing (NLP) task and will be described in the next chapter.

4.2.1.4 Natural language processing

The NLP task was performed using the LingPipe library. The tool evaluation focussed on free open source software (FOSS) frameworks. Commercial products, such as the IBM SPSS suite, were not evaluated. However, Amazon's cloud based ML service was also evaluated. The following NLP frameworks were evaluated:

LingPipe, Mallet, WEKA, RapidMiner, NLTK and Amazon ML

Some NLP tools, such as LingPipe and RapidMiner, have a dual-licensing model. They offer a free basic license and a commercial license offering additional features or services. There are many other frameworks such as OpenNLP, Googles Speech API or GATE, to name a few.

LingPipe was chosen for several reasons:

- LingPipe is an NLP toolkit but also has implementations of ML algorithms such as classification and clustering schemes

- LingPipe has performed very well for all NLP and classification tasks that were executed
- LingPipe is written in Java and has a native Java API
- LingPipe has a very complete documentation, a shortcoming of some of the other toolkits
- LingPipe is widely used in academia and the industry; a Google Scholar search returns more than 2000 publications.
- LingPipe is Big Data ready
- LingPipe 1.0 was released in 2003 (Baldwin & Dayanidhi 2014) and has matured since.

For this study version 4.1 was used. LingPipe is a very popular NLP framework in academia and has been used in many studies (Lloret et al. 2012; Asur & Huberman 2010; Pang, B & Lee, L 2008; Carpenter, B 2007; Carpenter, B 2004).

Word sense disambiguation determines the meaning of a polysemous word in a specific context. For instance, ‘apple’ can refer to the company Apple Inc., or the fruit of the deciduous apple tree. Word sense disambiguation was not necessary since the words ‘iPhone’ and ‘iPad’ have no other meaning than the devices by Apple Inc. An analysis of the query results, using human judgement, showed that less than 1% of the captured tweets were about the fruit, and not the company.

Two data pre-processing tasks were performed: basic subjectivity analysis to remove objective tweets, and data deduplication.

4.2.1.4.1 Data deduplication

Near duplicates were eliminated using the Jaccard index, also called the Jaccard similarity coefficient. This commonly used measure calculates the likelihood of a node that is a neighbour of either x or y to be a common neighbour. It can be formulated as the number of common neighbours divided by the total number of neighbours of either x or y (Zafarani, Abbasi & Liu 2014, p. 328):

$$\sigma(x,y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}$$

Equation 4-1: Jaccard index

LingPipe has an implementation of the Jaccard index.

The Jaccard index divides the intersection of tokens from the two strings over the union of tokens from both strings (Baldwin & Dayanidhi 2014, p. 119). Tokenization is the process of extracting useful words, phrases or symbols from text sequences. A tokenizer breaks the text into text sequences that are used to calculate the distance. LingPipe has several implementations of tokenizers. A standard approach of statistical modelling is n-gram analysis. It is based on counting frequencies of occurrences of short symbol sequences of length up to n (called n-grams) (LeCun, Bengio & Hinton 2015, p. 441). If the vocabulary size is *V*, the number of possible n-grams is in the order of V^n. LingPipe has an NGramTokenizer. N was set to 8. Instead of just splitting texts up into words, it splits texts up into sequences of 8 words. If texts are analysed at the single word level, sentiment polarity shifters might be ignored.

Data deduplication reduced the data set by an average of:

Query term	Percent
iPhone	8.381562
Apple	6.61280475
iPad	4.85566755

Table 4-3: Data deduplication percent per term

4.2.1.4.2 Basic subjectivity analysis

For basic subjectivity analysis, a statistical *n*-gram Language Model (LM) was used. Language models define probability distributions $p(\sigma)$ over strings $\sigma \in \Sigma^*$ drawn from a fixed alphabet of characters Σ (Carpenter 2007, p. 2). The LM classifier is a learner that uses categorized character sequences, not single isolated words as many ML schemes do with the bag-of-words, unigram approach. LM classifiers relax some of the independence assumptions of naïve Bayes, allowing a local Markov chain dependence in the observed variables, while still permitting efficient inference and learning (P eng, Schuurmans & Wang 2004, p. 317).

Training data is often created manually. Alternatively a body of annotated training data that has been made available on the Internet can be used. Here training data from Pang and Lee (2004) with a body of 10,000 records has been used for training an LM classifier. The Dynamic Language Model Classifier from LingPipe was used. A DynamicLMClassifier is a language model classifier that accepts training events of categorized character sequences (LingPipe API 2016). It is based on a multivariate estimator for the category distribution. *n* was tested for values from 4 to 11. The categories are “objective” and “subjective”, *n*=8 yielded the overall best results. Whereas the total accuracy, F1 and kappa statistic values, started to deteriorate for values above 8, the area under curve still improved. Table 4-4: Summary statistics for basic subjectivity analysis shows the summary statistics for the subjectivity analysis:

n-gram	Total accuracy	F1	kappa	PR	ROC
11	0.916	0.916	0.832	0.799797	0.791308
10	0.917	0.917	0.834	0.799184	0.791236
9	0.919	0.919	0.838	0.795737	0.789034
8	**0.921**	**0.921**	**0.842**	**0.79233**	**0.785358**
7	0.914	0.914	0.828	0.786075	0.780812
6	0.915	0.915	0.83	0.777449	0.77448
5	0.924	0.924	0.848	0.761533	0.76123
4	0.909	0.909	0.818	0.719389	0.725968

Table 4-4: Summary statistics for basic subjectivity analysis

4.2.1.5 Classifier evaluation

There are several metrics to evaluate a classifier. For this study the total accuracy, F1 score, kappa coefficient, precision and recall and the receiver operating characteristic was used. Traditionally, information retrieval evaluation of accuracy, precision, recall and a combined f-measure score have been used (Ting, Wu & Ho 2010). The most common metric for evaluating classifiers is the accuracy. The accuracy is also referred to as the overall recognition rate of the classifier, that is, it reflects how well the classifier recognizes tuples of the various classes (Han, J & Kamber, M 2006, p. 360). The accuracy values for an unbalanced data set task are usually skewed because of the disproportion between matches and non matches (Ting, Wu & Ho 2010). For this reason, other measures such as precision and recall values have also been used. A measure called the Kappa statistic takes the expected figure into account by deducting it from the predictor's successes and expressing the result as a proportion of the total for a perfect predictor (Witten, Frank & Hall 2011, p. 166). The maximum value of Kappa is 100%.

The result of each test is represented as confusion matrix. A confusion matrix is an unambiguous view of the classification accuracy. It is called confusion matrix because it

is easy to see which categories the learner confuses. For instance, a learner classifying text into British, Australian and American English would be expected to be highly confusable. Confusion matrices can also be used to compare multiclass classifiers. A perfect classifier has all zeroes except in the cells that are located diagonally from the top left to the bottom right. They represent the correctly classified instances.

A confusion matrix is a two dimensional contingency table with identical sets in both dimensions representing the instances in a predicted class. It is also called an error table. A confusion matrix can be used for binary and multiclass prediction with a row and column for each class. Confusion matrices can be used to quantitatively compare two classifiers over a fixed set of categories. This allows a more detailed analysis than just the proportion of correct guesses such as the accuracy. An ideal classifier has only values bigger than zero on the main diagonal. Table 4-5: Confusion matrices, shows the confusion matrices of the tests:

n=11	Objective	Subjective
Objective	453	47
Subjective	37	463

n=10	Objective	Subjective
Objective	455	45
Subjective	38	462

n=9	Objective	Subjective
Objective	457	43
Subjective	38	462

n=8	Objective	Subjective
Objective	458	42
Subjective	37	463

n=7	Objective	Subjective
Objective	459	41
Subjective	45	455

n=6	Objective	Subjective
Objective	458	42
Subjective	43	457

n=5	Objective	Subjective
Objective	462	38
Subjective	38	462

n=4	Objective	Subjective
Objective	451	49
Subjective	42	458

Table 4-5: Confusion matrices

The total accuracy, also called confidence, is expressed as the proportion of the correctly predicted instances from all instances. It can be expressed as percentage by multiplying it with 100. It is used to decide whether a model is good enough to make robust predictions. Accuracy alone is usually not enough to make these predictions.

The F_1 score, also F-score or F-measure measures the test accuracy of a binary classification. It is defined as:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Equation 4-2: F1 score

The F-score balances precision and recall.

Precision is the number of true positives, divided by the number of true positives and false positives. It is also called Positive Predictive Value (PPV). It is a measure for a classifier's exactness. Recall is the number of true positives divided by the number of true positives and false negatives. It is also called sensitivity or True Positive Rate (TPR). It is a measure of a classifier's completeness.

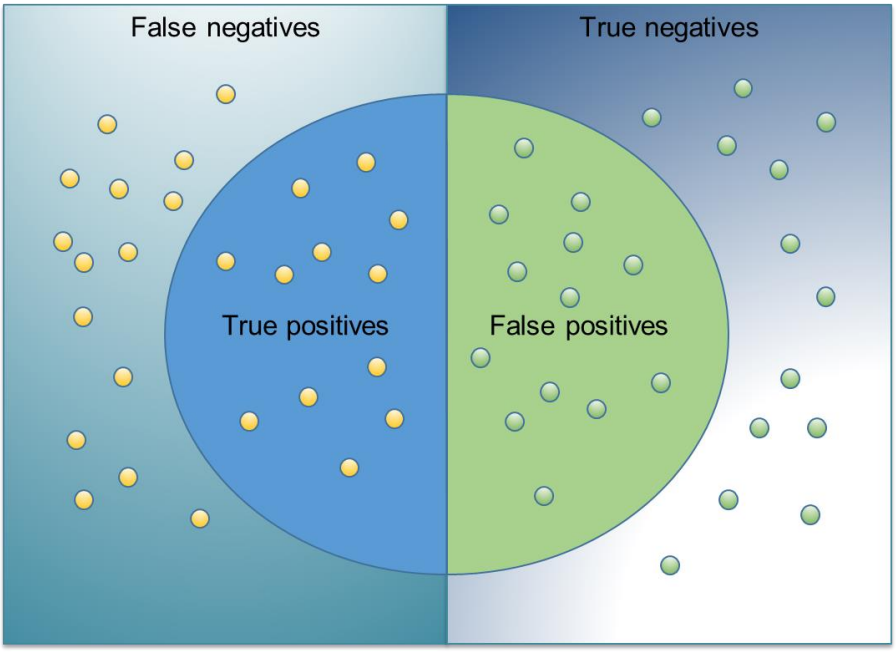


Figure 4-5: Precision Recall

Precision is defined as:

$$precision = \frac{tp}{tp + fp}$$

Equation 4-3: Precision

where *tp* = *true positives* and *fp* = *false positives*.

Recall is defined as

$$recall = \frac{tp}{tp + fn}$$

Equation 4-4: Recall

Where *tp* = *true positives* and *fn* = *false negatives*.

The Kappa statistic is used to measure the agreement between predicted and observed categorizations of a dataset, while correcting for an agreement that occurs by chance (Witten, Frank & Hall 2011, p. 166). This measure takes into account the probability of

random agreement between the predicted and the actual observed values and it is computed as:

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

Equation 4-5: Kappa coefficient

Where:

$$P(a) = \frac{m_{11} + m_{12}}{m}$$

Equation 4-6: Observer agreement

is the observed agreement and *P(e)* is the probability of random agreement, that is, the probability that the actual and the predicted coincide assuming independence between predictions and actual values (Arias, Arratia & Xuriguera 2014, p. 10). *m* is the total number of predicted values, m_{11} is the number of correct predictions (or hits) for the upward movement, m_{12} is the number of failed predictions (or misses). κ is one when the predicted and actual values agree.

Receiver Operating Characteristic (ROC) curves depict the performance of a classifier without regard to class distribution or error costs. They plot the true positive rate on the vertical axis against the true negative rate on the horizontal axis (Witten, Frank & Hall 2011, p. 172). The top left corner is the “ideal” point where the false positive rate is zero and the true positive rate is one. Hence, the larger the Area Under Curve (AUC) the better the classifier performs.

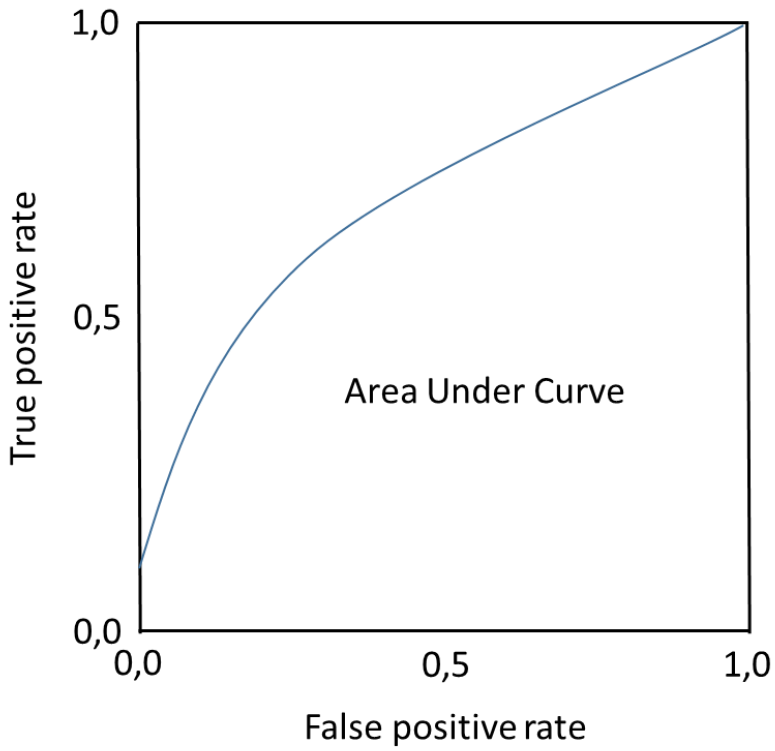


Figure 4-6: Receiver Operating Characteristic

ROC curves are used to compare the performance of binary classifiers. In real-world settings, ROC curves are usually not as smooth as the one shown in Table 4-6, but are more stepped.

An average of 40 % of the records was eliminated as a result of the subjectivity analysis. Table 4-6: Percentage of objective statements shows the percentages per month and search term:

	iPhone	iPad	Apple
November	41.993893	41.32689	39.33967
December	41.89642	41.90201	39.47344
Total %	41.9451565	41.61445	39.40656

Table 4-6: Percentage of objective statements

The accuracy of every step was always verified using human judgement. Human judgement is a manual verification process where randomly selected datasets are inspected to get an estimate for the performance of the classifier. For instance, tweets that were removed for containing objective statements, not opinions, were randomly selected to substantiate that they are not actually opinionated.

4.2.2 Predictive analysis phase

After the data had been pre-processed, the data analysis was performed using ML techniques. As stated before in chapter 1.3 Research methodology, there are many different ML techniques. Classifying tweets into positive and negative tweets is a binary classification problem and supervised ML techniques apply. Several learning schemes were trained and compared.

Classifiers should make no mistakes. For instance, a spam filter that classifies emails into legitimate and spam should always correctly detect spam. However, in real-life problems there are often a number of erroneously classified records. The number of errors should be small, so the best performing classifier was selected for SA. To compare them, a classifier needs a measure for the confidence of the classification. This is usually a score or a probability such as Precision Recall (PR) or the F-score.

First, the learning schemes have to be trained. To train the models, several corpuses of hand-classified tweets were evaluated. Labelled data is often called truth data, ground truth or golden standard data. It is expensive to produce in any quantity and the cleanest articulation of what is being done (Baldwin & Dayanidhi 2014, p. 82). Gold data is often labelled manually or verified manually. Labelling is also called annotating. The training data was collected over a one-week period. A day's data is likely to be correlated, for

instance to an announcement or another event and thus biased. That is why data over a period of one week was collected. The quality of the test set was verified using human judgement. Some records were randomly selected and manually verified.

The classifiers were trained and evaluated using n-fold cross-validation. The training data was divided into n=10 sets or folds. 9 folds were used for training, 1, the holdout set, for testing. This process is also called holdout cross-validation. Extensive tests on numerous different datasets, with different learning techniques, have shown that 10 is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up (Witten, Frank & Hall 2011, p. 153).

To improve cross-validation performance, the folds were verified to make sure the class label was properly represented in each fold, otherwise the learner is poorly trained. This process is called *stratification* and the whole process *stratified holdout cross-validation*. The folds were randomly selected and the whole process including random sampling was repeated 10 times to mitigate the bias. The error rates of each iteration were averaged over all iterations. The process was repeated 9 times until all 10 folds had been used once for testing.

The workflow used to improve the performance of the learner was as follows:

1. Verify the cross-validation performance
2. Verify the error
3. Analyse the errors by looking at the wrongly classified datasets
4. Adjust the system
5. Evaluate again

Learning models that improved performance during cross-validation also increased performance on new data.

Cross-validation did not reliably predict the performance on new data. As a consequence, four different error categories were evaluated:

For a given category X:

True positive: The classifier guessed X, and the true category is X

False positive: The classifier guessed X, but the true category is a category that is different from X

True negative: The classifier guessed a category that is different from X, and the true category is different from X

False negative: The classifier guessed a category different from X, but the true category is X (Baldwin & Dayanidhi 2014, p. 109)

These error categories were used to determine the evaluation metrics *Precision Recall* (PR) and *Specificity*.

Precision is defined as: true positive / (false positive + true positive)

Recall is defined as: true positive / (false negative + true positive)

Specificity is defined as: true negative / (true negative + false positive)

It has to be noted that classifications are not mutually exclusive; a tweet can have a positive and a negative statement and thus belong to both classes.

Following models were trained and evaluated:

- Perceptron and multilayer perceptron classifier
- Decision tree induction
- Multigram language model
- Naïve Bayes
- Logistic regression classifiers

Only the Language model, naïve Bayes and Logistic Regression, was used for the actual analysis. The perceptron and decision tree classifiers did not perform well and were not used for analysis. All classifiers tested were using tokenized input. Tokenization breaks a stream of text into words, n-grams, phrases, phonemes or other meaningful elements. They are called tokens and are used for document representation. LingPipe's classification, tagging, and entity extraction are all based on n-gram character language models (Carpenter 2007, p. 2). Document representation refers to the selection of appropriate features to represent documents (Shen et al. 2006, p. 672). These feature vectors serve as input for the classification algorithms.

4.2.2.1 Multigram language model

LingPipe's classification, tagging, and entity extraction are all based on n-gram character language models (Carpenter 2007, p. 2). A language model (LM) is a probability distribution over a sequence of words. It assigns a probability to the character sequence, the n-gram or multigram, and then calculates the likelihood that it belongs to a certain category. LingPipe adopts a standard random processing approach to n-gram language models, where probabilities are normalized over strings of a fixed length (Carpenter 2007, p. 2). The LM classifier accepts categorized character sequences as learning input and the multigram size. The best performance was obtained using 8-gram character sequences. It is a supervised learning scheme. The estimator for each category is initialized with a uniform Dirichlet prior with α=1 and using Laplace smoothing. Twitter users, and SM users in general, tend to use slang or orthographically incorrect words such as *"I looooooooove my new car"* to emphasise the sentiment expressed in the tweet, or *"I luv this restaurant"* to be cool. These words are likely to not be seen in the

test data. Smoothing allows assigning non-zero priors to words not seen in the sample and thus mitigates the problem of finding orthographically incorrectly spelled words. LM classifiers can be used for active learning in a tag-a-little, learn-a-little learning mode. The same learner was used for basic subjectivity analysis but with a different training set and with boundary character n-gram models. With an order of 8, the LM classifier will use a sentence *"I love my new car"* and produce "I", "I ", "I l", "I lo" training instances with up to maximum 8 characters.

First, the trained learner was evaluated manually and the overall performance of the manual tests was not satisfactory. Table 4-7: Test sentences shows some of the test sentences used to manually verify the learner:

Test sentences
I love the new iPhone
this mattress had a valley after two months
the new iPod is very user friendly
the iPad is very easy to use
the iPad is excellent
the iPhone is great
my car cost me an arm and a leg
why would anyone buy an iPhone
the new Tesla isn't great
this beer is flat

Table 4-7: Test sentences

The test sentences used phrases from other domains, such as beverages, cars and mattresses, to verify how well the model generalises. Also, manual testing is used to verify how well smoothing works. Using a learner trained with the smaller corpus of 737 tweets, some probabilities of test sentences are shown in Table 4-8:

Sentence	Probabilities
I love the new iPhone	positive 1.00
	negative 0.00
this mattress had a valley after two months	negative 1.00

	positive 0.03
the new iPod is very user friendly	negative 0.98
	positive 0.02
the iPad is very easy to use	negative 0.53
	positive 0.47
the iPad is excellent	positive 1.00
	negative 0.00
the iPhone is great	positive 1.00
	negative 0.00
my car cost me an arm and a leg	negative 1.00
	positive 0.00
why would anyone buy an iPhone	negative 1.00
	positive 0.00
the new tesla isn't great	negative 0.70
	positive 0.30
this beer is flat	negative 0.57
	positive 0.43

Table 4-8: Test sentence results with 737 training tweets

Two of the wrongly classified sentences, *"the new iPod is very user friendly"*, and *"the iPad is very easy to use"*, are from the domain of opinions from the test dataset; hence the manual tests showed that the performance of the learner is not satisfactory. Using the larger dataset of 1,048,575 tweets, the manual tests showed an even poorer result. The learner was then evaluated using common statistical values such as the total accuracy, the F1-score, and kappa.

Table 4-9: Summary statistics for Language Model classifier shows the summary results of the evaluation of the trained classifier:

n	Total accuracy	F1	kappa
737	0.986486	0.9864865	0.972973
1048575	0.982051	0.9820511	0.964102

Table 4-9: Summary statistics for Language Model classifier

The confusion matrices are shown in in Table 4-7:

n=74	Objective	Subjective
Objective	36	1
Subjective	0	37

n=51758	Objective	Subjective
Objective	26739	776
Subjective	153	24090

Table 4-10: Confusion matrices

Contrary to the manual tests, these statistics show a very high accuracy and predictive performance. In conclusion, the LM classifier is not a suitable learner for the classification task needed for this study.

4.2.2.2 Naïve Bayes

The Naïve Bayes classifier is one of the most popular classifiers (Baldwin & Dayanidhi 2014, p. 190). It is used for example in spam filters (Tretyakov 2004). The word *naïve* refers to the fact that the features are assumed to be independent. This is a naïve assumption especially for texts where words are dependent on each other to form meaningful sentences. Nevertheless, it is used frequently in practice as it is relatively simple and works well for many text classification problems. The Naïve Bayes classifier has the following characteristics:

- It tokenizes character sequences into words with frequencies called bag of words. This particular form is called the multinomial Naïve Bayes classifier and the word order is immaterial
- Each word is a Boolean attribute
- It requires two or more categories that are exhaustive and mutually exclusive
- It is configurable for various kinds of unknown token models

The naïve Bayes classifier calculates the probability of obtaining word *i* from all the documents in category *H*. It assumes the probability is independent of the words context

and position. A token-based naïve Bayes classifier computes the joint token count and category probabilities as follows (Baldwin & Dayanidhi 2014, p. 191):

$$\Pr(T,H) = P(T \mid H) \cdot P(H)$$

Equation 4-7: Category probability

where *T* are the tokens. Each token is assumed independent, and the probability of all tokens is the product of the probability of each token. These probabilities are used to calculate the maximum likelihood estimate for the model. Since tokens that were unseen during training result in a zero probability estimate, smoothing was used to mitigate the problem. A technique called *Laplace smoothing* was used. In the standard Naive Bayes approach, Laplace smoothing is commonly used to avoid zero probability estimates (Peng, Schuurmans & Wang 2004, p. 318). Laplace smoothing is also called *additive smoothing* and it allows adding probabilities to words that do not appear in the training data. Experiments show that smoothing substantially increases the accuracy of predictions (Witten, Frank & Hall 2011, p. 252).

LingPipe has several implementations of the naïve Bayes classifier. The classifier was first trained using tweet737 manually annotated tweets. The probabilities are listed in Table 4-8:

Sentence	Probabilities
I love the new iPhone	positive 0.96
	negative 0.04
this mattress had a valley after two months	negative 0.97
	positive 0.03
the new iPod is very user friendly	positive 0.62
	negative 0.38
the iPad is very easy to use	negative 0.63
	positive 0.37
the iPad is excellent	positive 0.53
	negative 0.47
the iPhone is great	positive 0.55

	negative 0.45
my car cost me an arm and a leg	negative 0.78
	positive 0.22
why would anyone buy an iPhone	negative 0.99
	positive 0.01
the new tesla isn't great	negative 0.74
	positive 0.26
this beer is flat	negative 0.85
	positive 0.15

Table 4-11: Test sentence results with 737 training tweets

With one exception, all tweets were correctly classified. A first analysis revealed that some of the seemingly more obvious sentences to classify were categorized with a lower probability whereas sentences more difficult to classify were categorized with a higher probability. For instance, *"the ipad is excellent"* was classified as positive with a probability of 0.53, whereas *"my car cost me an arm and a leg"* was classified as negative with a probability of 0.78.

A second dataset of 1,048,575 annotated tweets was used for training, 554.477 positive and 494,098 negative from Sentiment140 (Sentiment140 2016). Other datasets that were tested for training contained, for instance, opinions about Apple, Google and Microsoft. However, they showed low accuracies during training and were not used again.

The pseudo code in Figure 4-7: Naive Bayes classifier shows the NB algorithm, where *C* is the class labels, *D* is the Data set, the corpus of tweets, and *t* is a term in a tweet:

```
TRAINNB(C, D)
   1. V ← EXTRACTTWEETS(T)
   2. N ← COUNTTWEETS(T)
   3. for each c ∈ C
   4. do N_c ← COUNTTWEETSINCLASS(D,c)
   5.        prior[c] ← N_c / N
   6.        text_c ← CONCATENATETWEETSOFALLTWEETSINCLASS(D,c)
   7.        for each t ∈ V
   8.        do T_ct ←  COUNTTOKENSOFTERM(text_ct, t))
   9.        for each t ∈ V
```

```
   10.       do countprob[t] [c] ← (T_cT+1) / Σ_t'(T_ct'+1)
   11. return V, prior, condprob

APPLYNB(C, V, prior, condprob, d)
   1. W ← EXTRACTTOKENSFROMDOC(V, D)
   2. for each c ∈ C
   3. do score[c] ← log prior[c]
   4.       for each t ∈ W
   5.       do score[c] += log condprob[t][c]
   6. return arg max_c ∈ C score[c]
```

Figure 4-7: Naive Bayes classifier

The smoothing variable did not have a considerable influence on the performance of the classifier, so a uniform model was used.

Table 4-12 shows the summary results of the evaluation of the trained classifier:

n	Total accuracy	F1	kappa
737	0.905405405	0.905405	0.810811
1048575	0.778874763	0.778875	0.55775

Table 4-12: Naïve Bayes classifier

The confusion matrices are shown in Table 4-13:

n=74	Objective	Subjective
Objective	33	4
Subjective	3	34

n=51758	Objective	Subjective
Objective	18628	8887
Subjective	2558	21685

Table 4-13: Confusion matrices

The results for the n=737 truth data was dramatically improved using cross-validation. The performance of the large, n=1048575 data set was deteriorated for all measured statistics. The model was seriously overfitted due to a truth data set that contained many non-relevant opinions. The result was a classifier that captures noise instead of pertinent data points.

Table 4-14 shows the summary results of the evaluation of the trained classifier using 10-fold cross-validation:

n	Total accuracy	F1	kappa
737	0.972972973	0.972973	0.945946
1048575	0.573611809	0.573612	0.147224

Table 4-14: Naïve Bayes classifier using 10-fold cross-validation

The confusion matrices are shown in Table 4-15:

n=74	Objective	Subjective
Objective	36	1
Subjective	1	36

n=51758	Objective	Subjective
Objective	13400	14115
Subjective	7954	16289

Table 4-15: Confusion matrices

As stated earlier in Section 4.2.2 having more than 10 folds does not improve the performance of the classifier. This was confirmed by training and evaluating the classifier with folds from 5 to 10.

An important factor in cross-validation is the proper representation of the classes in both the training data and test data. In other words, classes should not be over- or underrepresented. In the worst case scenario, a class which is not represented in the training data could be overrepresented in the test data. Therefore, the random sampling has to be done in a way such that each class is well represented in both datasets, a process called stratification. However, stratification is only a weak safeguard against uneven representation of instances. A more general way to mitigate any bias caused by the particular sample chosen for holdout is to repeat the whole process by training and testing, several times, with different random samples (Witten, Frank & Hall 2011, p. 152). It is for this reason that the corpus was randomly permutated using a randomizer.

4.2.2.3 Logistic regression

Logistic regression (LR) classifiers are discriminative probabilistic classification models. They use feature vectors extracted from the corpus by a chain of conditional random fields (CRF)-specific feature extractors. CRF-based methods can obtain an improvement in accuracy on long texts compared to some existing rule-based or supervised learning methods (Zhang 2013, p. 19). They belong to the class of linear classifiers since the output is a linear expression for each class. This scheme is sometimes called multiresponse linear regression (Witten, Frank & Hall 2011, p. 125). LR uses real-valued feature vectors with weights as input and applies a vector of coefficients. LR feature extraction is not limited to characters or tokens It can use unlimited feature extraction which allows for arbitrary observations. Logistic regression is one of the best probabilistic classifiers, measured in both log loss and first-best classification accuracy across a number of tasks (LingPipe 2015). It almost certainly is one of the best performing classifiers available, albeit at the cost of slow training and considerable complexity in configuration and tuning (Baldwin & Dayanidhi 2014, p. 202). LR obey the maximum entropy paradigm which states that the correct distribution *p(a, b)* is that which maximizes entropy, or “uncertainty”, subject to the constraints. The contraints represent “evidence”, i.e. the facts known to the experimenter (Ratnaparkhi 1997, p. 2) where *p* is the probability of class *a* occurring in context *b*. Hence a LR classifier is a maximum entropy classifier.

LR classifiers are highly customizable. The training underwent many iterations with different parameters and feature extraction methods until the classifier performed at its optimum. The methods and parameters are described in the next section.

The TokenFeatureExtractor was used to extract the tokens during construction to create the features. A minimum feature count was used since LR tends to overfit on low token count features that exist by chance in a training data set. The addInterceptFeature defines whether a category feature exists. If a category is very rare or very common it should be captured. The noninformativeIntercept defines how the feature is handled. If true, no priors are applied to the intercept. Priors help to prevent LR from overfitting by pushing coefficients towards 0. RegressionPrior is the expected variance of the features. Low variance will push coefficients more aggressively to zero. Priors, in this context, function as a way to not be over-confident with observations about the world (Baldwin & Dayanidhi 2014, p. 205). Simulated annealing is an optimization method. The AnnealingSchedule instance calculates the learning rate for a specific epoch. An epoch is a learning iteration.

The LR classifier was trained in several configurations. The first training was with a tokenizer feature extractor with and without cross-validation. The results for both truth datasets are summarized in Table 4-16 and Table 4-17:

LR classifier without cross-validation:

n	Total accuracy	F1	kappa
737	0.918918919	0.918918919	0.837837838
1048575	0.837880134	0.837880134	0.675760269

Table 4-16: Logistic regression without cross-validation

n=74	Objective	Subjective
Objective	33	4
Subjective	2	35

n=51758	Objective	Subjective
Objective	22158	5357
Subjective	3034	21209

Table 4-17: Confusion matrices

LR classifier with 10-fold cross-validation:

n	Total accuracy	F1	kappa
737	0.905405405	0.905405405	0.810810811
1048575	0.833861432	0.833861432	0.667722864

Table 4-18: Logistic regression with 10-fold cross-validation

n=74	Objective	Subjective
Objective	33	4
Subjective	3	34

n=51758	Objective	Subjective
Objective	22039	5476
Subjective	3123	21120

Table 4-19: Confusion matrices

LR classifier with character n-gram:

n	Total accuracy	F1	kappa
737	0.959459459	0.959459459	0.918918919
1048575	0.878453572	0.878453572	0.756907145

Table 4-20: Logistic regression using character n-grams

n=74	Objective	Subjective
Objective	36	1
Subjective	2	35

n=51758	Objective	Subjective
Objective	23558	3957
Subjective	2334	21909

Table 4-21: Confusion matrices

Counter intuitively the feature extraction approach using n-grams performed significantly better than the words/tokens or stemmed words approach. Smaller datasets usually benefit from lower order (number of tokens) n-grams (Baldwin & Dayanidhi 2014, p. 187). Cross-validation did not improve the classification accuracy but significantly slowed down training, so the cross-validation approach was not further pursued.

4.2.2.4 Summary

Based on the results, the Naïve Bayes classifier, trained on the Apple specific dataset and with 10-fold cross-validation, and the LR classifier, trained on the same dataset with character n-grams, were used for classifying the collected tweets. They scored the highest in terms of accuracy, F1 score and kappa coefficient as well as manually verifying the classifiers with sample opinions.

Two of the other trained classifiers, the LM classifier and the perceptron, did not yield satisfactory results and were nor used for classification. Following result sets were used for correlation analysis: Naive Bayes with 10-fold cross-validation and LR with character n-gram. For each classifier, the total number of positive and negative tweets, the total number of positive and negative deduplicated tweets and the total number of positive and negative tweets, after subjectivity analysis, were determined and used for correlation analysis.

4.2.2.5 Granger causality test

The Granger causality test is an econometric technique. It is used to determine if one time series has predictive information for another (Wlodarczak, Soar & Ally 2015, p. 4). It uses different lags of one series to model changes in the second time series. The Granger causality test for two scalar-valued, stationary, and ergodic time series $\{X_t\}$ and $\{Y_t\}$ is defined as:

$$F(X_t \mid \mathrm{I}_{\mathrm{t-1}}) = F(X_t \mid (\mathrm{I}_{\mathrm{t-1}} - Y_{t-Ly}^{Ly})), t = 1,2...$$

Equation 4-8: Granger causality test

Where $F(X_t|I_{t-1})$ is the conditional probability distribution of X_t given the bivariate set I_{t-1} consisting of an L_x-length vector X_t and an L_y-length vector of Y_t (Wlodarczak et al. 2015, p. 3). The F-test, t-test or Wald test (used in R) are calculated to test the following null and alternate hypotheses:

H_0: $\alpha_i = 0$ for each i of the element [1,k]

H_1: $\alpha_i \neq 0$ for at least 1 i of the element [1,k]

where k is the number of lags in the time series. The F-test is most often used to compare statistical models in order to identify the model that is best fitted for the population from which the data has been sampled. Essentially, we are trying to determine whether X provides more statistical information about future values of Y than past values of Y alone. The null hypothesis generally refers to the fact that there is no relationship between two observations. It is important to notice that we are not trying to prove actual causation; we are only trying to prove that two values are related by some phenomenon.

The R library “lmtest” contains the Granges causality test procedures. It has to be loaded first in the R script. The R “grangertest” function takes a bivariate time series on L lags and a dataset as input. The result of the grangertest function is DF, the degree of freedom, that is the number of observations, the F-statistic and the corresponding probability, $Pr(>F)$, the p-value. If $Pr(>F) < \alpha$, where α is the desired level of significance, we reject the null hypothesis of no Granger causality.

Figure 4-8: Sample grangertest output shows a sample output of the grangertest function:

```
Granger causality test
```

```
Model 1: Open ~ Lags(Open, 1:1) + Lags(Apple, 1:1)
Model 2: Open ~ Lags(Open, 1:1)
  Res.Df Df      F Pr(>F)
1     38
2     39 -1 0.0191 0.8907
```

Figure 4-8: Sample grangertest output

α indicates the magnitude of relationship observed. A common value for p is 0.05. However, it is important to mention that a small p-value does not imply causality in a theoretical sense. Also, Granger causality is not a test for strict erogeneity. From a statistical point of view, p equal or less than 0.05 denotes statistical significance and significance is not related to causality. If the p-value is more than 0.05, we cannot reject the NULL hypothesis. If the p-value is less than 0.05, we can reject the NULL hypothesis. The 0.05 significance level is also called the cut off. However, the "statistical significance", interpreted as "$p \leq 0.05$", is not sufficient to support a scientific hypothesis. The NULL hypothesis test is a reductio ad absurdum, which in essence states that a claim is valid by demonstrating that a counter-claim is improbable. Credible Granger causality analysis appears to require post-sample inference, as it is well-known that in-sample fit can be a poor guide to actual forecasting effectiveness (Ashley & Tsang 2014, p. 72).

We also need to verify that *Y* does not provide any information about *X*, otherwise there is likely an exogenous variable *z* that is better suited for Granger causation.

A sample R plot of time series is shown in Figure 4-9: R time series plot of Apple tweet frequency and share price:

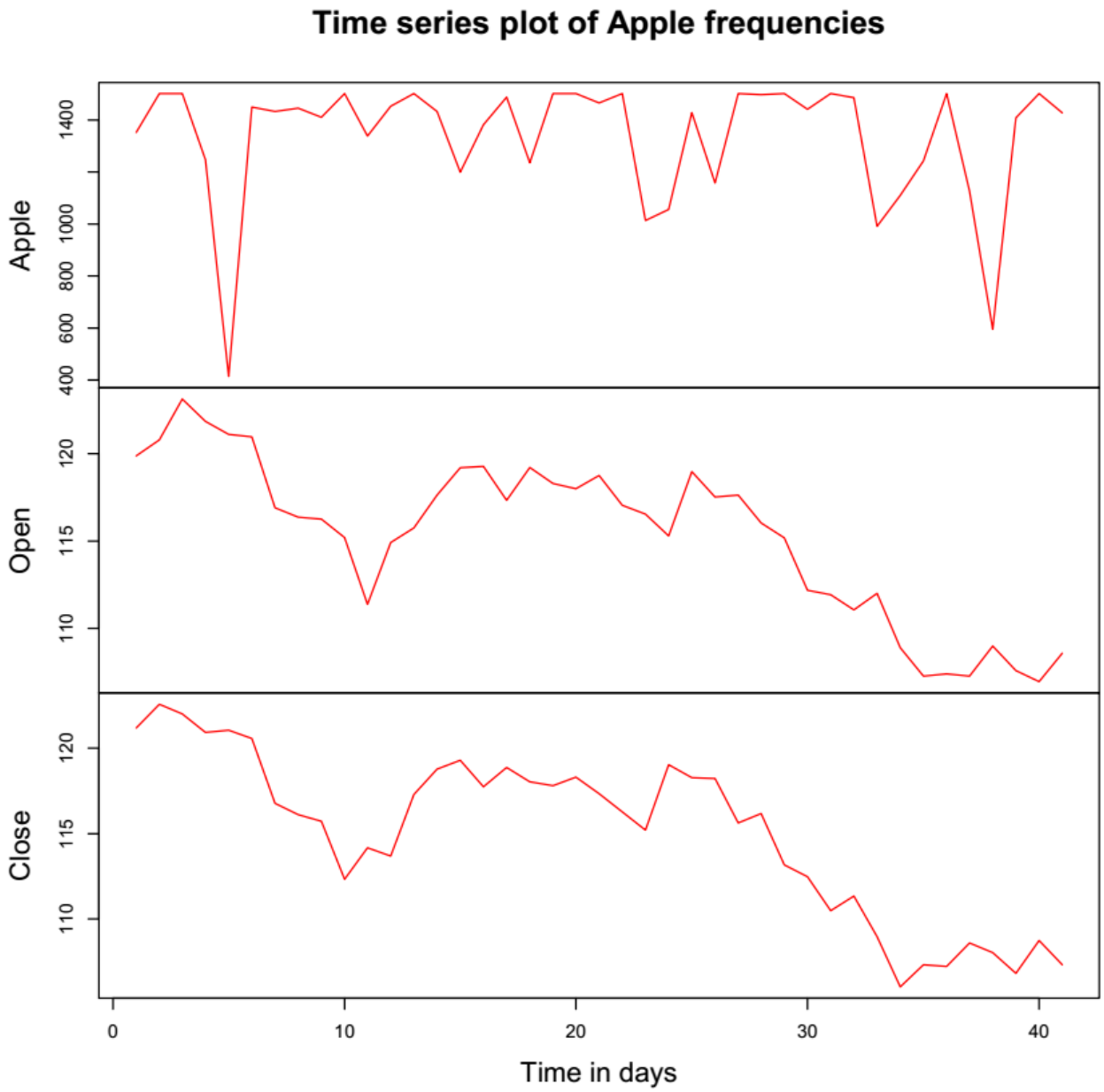


Figure 4-9: R time series plot of Apple tweet frequency and share price

Whereas the opening and closing quotes of the Apple share show a clear downwards trend, the Apple tweet frequency shows a ragged line with no clear trend. If Apple Granger causes the Open or Close quote time series, the patterns in Apple would approximately repeat in Open or Close after a time lag. In the above example, there is no

obvious pattern in the Apple time series that seems to repeat in the Open or Close time series so there is no Granger causality. This is also confirmed by the result, a probability of 0.8907, for example ~89%, which is considerably higher than 5%, so we cannot reject the NULL hypothesis. For this reason, the R script plots the time series for visual comparison and does not only rely on the p-value.

4.3 Results

For every data set, that is, the tweets collected over the month of November and December 2017, the frequencies of all collected data without classification and the classified data using naïve Bayes and Logistic regression was submitted to the Granger causality test. Since trading happens only on weekdays, tweets from the weekends were discarded. A total of 27 tests were performed. Only the lag that yielded the most significant results is listed.

The results are shown in following tables.

4.3.1 Apple datasets

The unclassified Apple data set, using only frequencies, showed for all three datasets, total number of Apple tweets, deduplicated Apple data set and the data set that underwent basic subjectivity analysis, values above 5% significance level, so we cannot reject the NULL hypothesis.

		Total	Deduplicated	Subjectivity Analysis
Apple	**Open**	89,07%	95,3%	93,83%
	Close	85,48%	89,48%	93,43%

Table 4-22: Apple unclassified

Open and Close in the second left column stands for open and close quote.

The datasets that were classified using the naïve Bayes and LR classifier showed significantly better results. The best significance levels were obtained using LR on the data set without deduplication and subjectivity analysis. With a value of 16,3% we still cannot reject the NULL hypothesis.

		Naïve Bayes			Logistic Regression		
		Total	Deduplicated	Subjectivity Analysis	Total	Deduplicated	Subjectivity Analysis
Apple	Pos open	34,75%	39,88%	30,79%	16,3%	32,74%	25,94%
	Pos close	54,11%	46,7%	53,13%	52,13%	24,83%	41,49%
	Neg open	37,6%	41,39%	37,6%	43,63%	42,72%	37,59%
	Neg close	46,56%	39,03%	41,29%	48,05%	45,41%	45,26%

Table 4-23: Apple classified

4.3.2 iPad datasets

The iPad datasets showed by far the lowest significance levels. Even the unclassified datasets showed significance levels of roughly more than 12%. Nevertheless the levels are still too high to determine a correlation and we cannot reject the NULL hypothesis.

		Total	Deduplicated	Subjectivity Analysis
iPad	Open	15,28%	12,63%	29,74%
	Close	16,72%	12,71%	26,98%

Table 4-24: iPad unclassified

The datasets using subjectivity analysis showed for both, the naïve Bayes and LR levels of slightly more than 7% and 8% respectively for the closing quote and negative tweets. This is still above the 5% benchmark. Considering that the datasets for positive tweets

and subjectivity analysis were above the 9% significance level for LR and 12% for the naïve Bayes classifier, there cannot be a correlation deduced and we cannot reject the NULL hypothesis.

		Naïve Bayes			Logistic Regression		
		Total	Deduplicated	Subjectivity Analysis	Total	Deduplicated	Subjectivity Analysis
iPad	Pos open	39,39%	13,55%	14,76%	37,23%	11,41%	9,892%
	Pos close	45,23%	11,57%	12,25%	42,57%	11,51%	9,115%
	Neg open	26,52%	9,955%	8,919%	24,38%	10,25%	10,75%
	Neg close	22,38%	9,167%	7,515%	18,34%	8,357%	8,314%

Table 4-25: iPad classified

4.3.3 iPhone datasets

The unclassified and classified iPhone datasets were all above 10%, the best value being 14,97% for unclassified tweets that went through subjectivity analysis and closing quotes, we cannot reject the NULL hypothesis for these datasets either.

		Total	Deduplicated	Subjectivity Analysis
iPhone	Open	87,43%	87,44%	49,79%
	Close	70,12%	52,21%	14,97%

Table 4-26: iPhone unclassified

		Naïve Bayes			Logistic Regression		
		Total	Deduplicated	Subjectivity Analysis	Total	Deduplicated	Subjectivity Analysis
iPhone	Pos open	81,7%	52,78%	49,8%	73,06%	48,64%	46,98%
	Pos close	53%	74,77%	75%	62,48%	75,04%	77,5%
	Neg open	60,61%	49,31%	54,21%	66,24%	53,06%	60,41%

	Neg close	42,47%	30,67%	29,02%	37,79%	27,23%	28,86%

Table 4-27: iPhone classified

In summary, only a weak correlation for some data iPad sets could be determined. Overall the significance levels were not sufficient to determine correlation and we cannot reject the NULL hypothesis.

4.4 Conclusions

As shown in the results section 4.3, there was a large span with the results, ranging from a significance of around 90% for the unclassified Apple datasets, to less than 10% for some iPad datasets. The variations were even substantial for datasets that underwent the same pre-processing steps. For instance, the unclassified, deduplicated Apple dataset produced significance levels of around 90%, whereas the unclassified, deduplicated iPad dataset showed levels of close to 13%. In the case of the iPhone datasets, there was a large difference in significance levels even for the opening and closing quotes.

Using unclassified data, using only frequencies did not produce any satisfactory results; the classification accuracies were below average. By far the best results were obtained using opinions on the iPad. Whereas data deduplication increased the accuracy of the Granger causality test considerably, basic subjectivity analysis did not improve the results significantly or even worsened it.

Opinion mining is highly domain specific. Different words for the same meaning are used depending on what is being reviewed, e. g. a movie, a car or a dish. This explains why a

smaller dataset with Apple specific opinions about the iPhone or the iPad performed considerably better than a much larger, general purpose dataset, even though the Apple corpus was much smaller. On the other hand, opinions are often expressed in similar ways using similar words within a domain. This finding contradicts the Big Data principle, where more data is always better than less data. Smaller datasets make training of learning models simpler since the features (observations) to be extracted are narrowed down.

Several learning models were trained for SA. Even simpler ones such as language model classifiers performed reasonably well if they were provided with enough training data. A major problem of language model classification is data scarcity, since most multigrams will not be seen during training. This can be mitigated using unigrams and making the assumption that the probability of a word only depends on the previous *n* words.

Cross-validation is a very effective training method, and in this study, it has dramatically improved the performance of classifiers when training is executed using slam datasets.

The NB classifier was the least prone to overfitting of the learners trained in this study. An overfit model is usually trained too close to the training data and does not generalize well. As a consequence, it captures noise or random error instead of the underlying relation.

Data pre-processing is at least as important as tuning the learner. In this study most of the time was consumed in the data pre-processing phase. Many different datasets were tested until satisfactory results were obtained. Again, this contradicted the Big Data paradigm where the dataset does not necessarily have to be of a high quality if there is sufficient data.

Twitter users can use the hashtag to highlight keywords in tweets since there is no subject line. For instance, a user could write "#iPhone" for a tweet about the iPhone. However, omitting the hashtag when using the Search API, did not appear to have a significant impact on the relevance of the collected tweets. Also, as an analysis using NVivo revealed, the majority of the analysed tweets did not contain hashtags. For instance, in a randomly selected and verified sample of 3197 Apple tweets over the period 15.11.2015 - 15.11.2015, only 301 tweets contained hashtags. Other datasets that were manually verified showed similarly low hashtag usage.

Contrary to the Big Data principle, which suggests that a large data set outdoes a small high quality sample, a dataset that contained only opinions about Apple products performed much better from a larger dataset with opinions on other companies such as Google and Microsoft.

Despite the fact that Apple sold more iPhones in 2015 than ever before (Pramuk 2015) this did not did not influence the share price as expected. It fell by 8.8% in the second quarter of that year. The main reason is probably that Apple, despite excellent results, did not met analysts' expectations (Higgins 2015).

This is also reflected in the fact that there are always considerably more tweets about the iPhone than, for instance, the iPad or Apple.

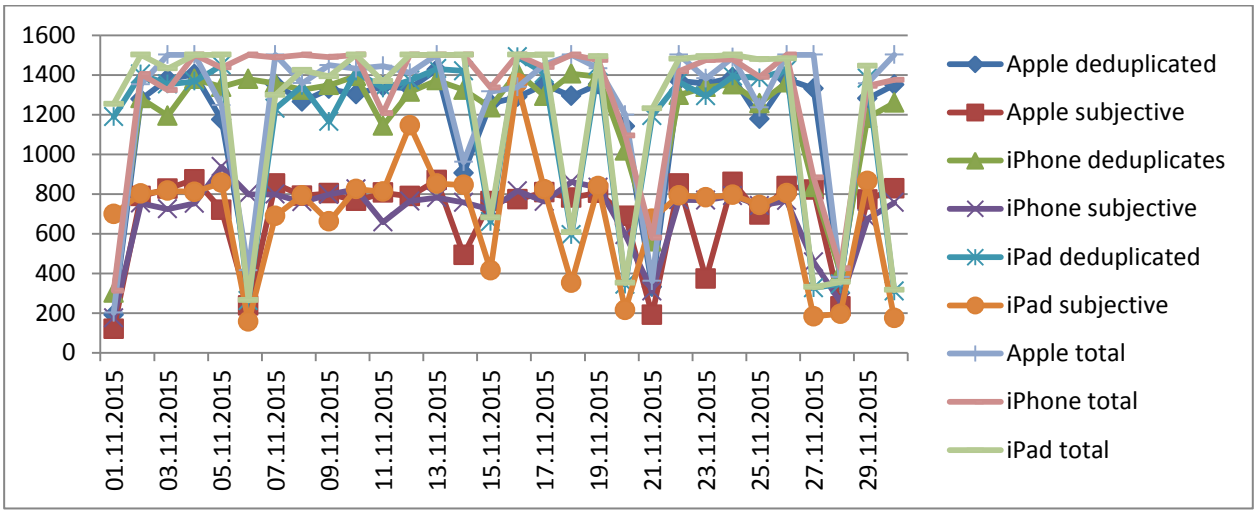


Figure 4-10: Frequencies of tweets

Both classifiers, the naive Bayes and Logistic Regression, performed about equally well. The term "Granger causality" is somewhat misleading, given that strong evidence of "Granger causality," may be weak evidence of "causality." Establishing causality, the relationship between cause and effect, is a difficult endeavour as there are likely to be many factors influencing the value of a share price. In fact, it is common to have a very accurate predictive ML model which nevertheless gives no information whatsoever about how or why something is happening (Huang et al. 2015).

Since the iPhone is by far the most important Apple product in terms of units sold and revenue generated (see Figure 4-11) for the time period when the data was collected, it is surprising to see that the best results were obtained using opinions on the iPad. It would have been expected, that it would also be the most significant for predicting the share price.

Apple Inc.
Q4 2015 Unaudited Summary Data
(Units in thousands, Revenue in millions)

	Q4 2015	Q3 2015	Q4 2014	Sequential Change	Year/Year Change
Operating Segments	**Revenue**	**Revenue**	**Revenue**	**Revenue**	**Revenue**
Americas	$21,773	$20,209	$19,750	8%	10%
Europe	10,577	10,342	10,350	2%	2%
Greater China	12,518	13,230	6,292	- 5%	99%
Japan	3,929	2,872	3,595	37%	9%
Rest of Asia Pacific	2,704	2,952	2,136	- 8%	27%
Total Apple	**$51,501**	**$49,605**	**$42,123**	**4%**	**22%**

	Q4 2015		Q3 2015		Q4 2014		Sequential Change		Year/Year Change	
Product Summary	**Units**	**Revenue**	**Units**	**Revenue**	**Units**	**Revenue**	**Units**	**Revenue**	**Units**	**Revenue**
iPhone (1)	48,046	$32,209	47,534	$31,368	39,272	$23,678	1%	3%	22%	36%
iPad (1)	9,883	4,276	10,931	4,538	12,316	5,316	- 10%	- 6%	- 20%	- 20%
Mac (1)	5,709	6,882	4,796	6,030	5,520	6,625	19%	14%	3%	4%
Services (2)		5,086		5,028		4,608		1%		10%
Other Products (1)(3)		3,048		2,641		1,896		15%		61%
Total Apple		**$51,501**		**$49,605**		**$42,123**		**4%**		**22%**

(1) Includes deferrals and amortization of related software upgrade rights and non-software services.
(2) Includes revenue from Internet Services, AppleCare, Apple Pay, licensing and other services.
(3) Includes sales of Apple TV, Apple Watch, Beats products, iPod and Apple-branded and third-party accessories.

Figure 4-11: Apple Q4 2015 results (Apple Inc. 2016)

The iPad time series provided the best results when the Granger test was applied. However, the predictive accuracy was still not high enough to show that the iPad time series could predict the Apple share price. In other words, using the methods applied in this study it was not able to determine any correlation between Twitter opinions and share price. **Based on these findings the research question as stated in Section 2.3 Research question and issues, cannot be affirmed through the use of the methods applied in this study and that other techniques and strategies might yield more definitive results.**

Also, from this study it is uncertain why the iPad series was more accurate from the iPhone time series.

The following possible reasons that may have influenced the result are offered here:

- There were not enough relevant tweets - including opinions from other SM sites might have provided different results
- The time frame during which the data was collected was too short
- The models used were not fit for purpose - other predictive models or correlation tests might produce different results
- No correlation between Twitter opinions and share price exists.

In future research, other approaches could be used. The Granger causality test assumes the analysed time series are covariance stationary. If the data is assumed non-stationary, windowing techniques could have been used, assuming that sufficiently short windows of non-stationary observations are locally stationary.

The Granger causality test is data driven; causal interactions are inferred directly from simultaneously recorded time series. Other possible approaches could be model driven, where a model is first elaborated and then assessed against the data, or based on VAR (Vector Auto Regression) such as the Toda-Yamamoto approach (Fan et al. 2013; Alimi & Ofonyelu 2013).

5 Conclusions and implications

5.1 Introduction

The previous chapters 3 Research methodology and 4 Data analysis elaborated on the research methodologies and the results of the study. The purpose of this chapter is to describe the contributions and potential implications of this research on the different bodies of knowledge.

5.2 Conclusions about the research problem

In using the framework developed in this research, namely by applying ML techniques and the Granger causality test, only limited predictive capabilities could be determined. There was a weak correlation between the Twitter moods and share price using mood states about the iPad: 7,515% using the naïve Bayes classifier and 8,314% using Logistic Regression, however only for closing quote and negative tweets, as shown in chapter 4.3 Results. This is above the commonly used 5% threshold. It is important to notice that a p-value equal to or less than 0.05 (5%) is a measure for significance, but not for causality. A low p-value merely supports the believed causal relationship. It calculates the probability under the assumption that the Null hypothesis is true. The value of 0.05 should not be considered an absolute value. Causality can create a significant result, but significance does not prove causality. Although correlation does not imply causation, correlation can be used as a hint to identify causality between random variables (de Siqueira Santos et al. 2013, p. 11). To establish causality, a randomized controlled trial is needed, causal graphs or propensity score matching techniques must

be applied. However, determining significance and properly establishing causality is beyond the scope of this study and can be a subject for future research.

The least accurate results were obtained using Apple moods except for one result: LR without data deduplication and basic subjectivity analysis provided a p-value of 0,163. Otherwise deduplication and subjectivity analysis mostly improved the results. Data pre-processing can improve the predictive performance of learners and a purified data set usually yields better results. This confirms that data pre-processing is at least as important as training and configuring the most suited classifiers.

Contrary to what would have been expected, the iPhone did not provide the best results. Also, pre-processing iPhone data sometimes worsened the outcome.

The performances of NB and LR were insignificantly different. Both classifiers provided high accuracies. However, LR classifiers are highly configurable, and more adjustments might have improved the classification performance.

To improve the result, several possible measures could be adopted:

- More training data could be collected
- Twitters streaming API could be used instead of the search API
- Other pre-processing steps such as outlier removal or normalization could improve the predictive performance
- Other Apple products such as the MacBook or iPod could be considered
- All opinions on any Apple product could be collected
- Other data analysis algorithms could be evaluated
- Other data sources could be included, for instance other SM sites or tweets in other languages

- Separate classifiers for positive and not positive, and negative and not negative classification with separate training data

In conclusion, using the experiments conducted in this study, no conclusive correlation could be detected, but the opportunities exist for more detailed analyses as suggested above.

5.3 Implications for theory

Contrary to the Big Data principle, where larger datasets produces better results from relatively small samples (Mayer-Schonberger & Cukier 2013), in this study, training datasets that were specifically tailored for the problem at hand performed better than much larger, but more general datasets. As described in Section3.1.2 Predictive analysis phase, the NB and LR classifiers had a lower accuracy with using the dataset with more than 1 million records. This was the case in all experiments performed in this study, independently whether they were trained with or without cross-validation. Also, using a dataset that had undergone data pre-processing such as data deduplication resulted in mostly higher accuracy results. Noisy examples are very likely to be misclassified, and so the set of stored exemplars tends to accumulate those that are least useful (Witten, Frank & Hall 2011, p. 245). In the worst case, the predictive model is constructed using noise, for instance random data that has been erroneously collected. If a parameter is totally random, then it cannot tell you anything meaningful about the data object and you can drop the parameter (Berman 2013, p. 135). This suggests that more data is not necessarily better from having a smaller, high-quality sample. In this study, datasets that were put through pre-processing steps such as deduplication and subjectivity analysis performed better most of the time, the learning cycles were shorter and the classifiers

were more accurate. However, there are many different algorithms and techniques for Big Data analysis. Depending on the method used, classification accuracy could be better if the data volumes are large despite noise in the dataset. Interestingly enough, it has been shown that when artificial noise is added to attributes (rather than added to classes), test-set performance is improved if the same noise is added in the same way to the training set (Witten, Frank & Hall 2011, p. 332). In this study, both, the NB and LR classifiers had a higher classification accuracy on a smaller, high quality data set. Also, they were less prone to overfitting. NB seemed to be more robust whereas LR was sometimes overfitted, depending on the parameterisation. Most learning algorithms try to learn from noisy data by modelling the maximum likelihood output or least squared error, assuming that noise effects average out (Schmidt & Lipson 2007, p. 1). However, if the noise distributions are not symmetrical in the datasets, this approach does not hold.

The same observation applies to cross-validation. Cross-validation is a highly effective training method (Witten, Frank & Hall 2011, p. 152). However, in this study cross-validation sometimes overfitted the learning scheme and a training cycle without cross-validation performed better. This is most likely due to noise or uneven distributions of the class label.

The results varied significantly depending on what data pre-processing methods were adopted. The most notable differences were the variability of the iPhone open and close quote datasets. Since the differences of the quotes for open and close values are small, as would be expected since no trading happens between the closing and opening of the stock market, the distributions in significance of the same datasets were surprising. Explanations for the variance are uneven noise or class label distributions, but also inconsistencies in the datasets. In circumstances where big data are produced, acquired,

aggregated, transformed, or represented, inconsistencies invariably find their way into large datasets (Du 2013, p. 64). Since the same methods for data collection, pre-processing and analysis were applied, the same inconsistencies should manifest themselves in all the data, and not just the iPhone datasets. Since much more tweets about the iPhone are posted in the period when the data was collected, the inconsistencies might have augmented due to the larger datasets. The inconsistencies could have been coupled with nonlinear components. In other words symmetric internal noise can be scaled, offset, and in general transformed to produce non-symmetric noise distributions on the output (Schmidt & Lipson 2007, p. 1). In this situation the inconsistencies have deformed the maximum-likelihood output, and the regressed models may no longer describe the analytical structure of the system. The inconsistencies might be off topic tweets that were erroneously collected, uneven distributions of class labels or noise as stated before, or overfitted learners. To overcome these issues, applying more or different pre-processing steps such as filling in missing data (Perera et al. 2014), TF-IDF (Term Frequency-Inverse Document Frequency) (Ting, Wu & Ho 2010), active learning (Vesdapunt, Bellare & Dalvi 2014) or clustering techniques (Zhong et al. 2016) might mitigate the issues.

In summary, adopting data pre-processing steps seemed to improve the results. However, considering the variability of some of the results, the question of whether a correlation exists between twitter opinions and share price cannot be conclusive.

5.4 Implications for policy and practise

ML algorithms are domain independent, whereas data pre-processing is highly domain specific. ML models are commonly used across many different fields in computing and engineering (Perera et al. 2014, p. 433). Data pre-processing depends on the type of data that is being mined. For instance, for analysing text, NLP techniques have to be applied. Selecting NLP techniques depend on what is being mined for. For example, if text is mined to interpret historic data or to make decisions to influence events that are likely to happen in the future. As we have seen in this study, data representation in the form of feature vectors is crucial for data analysis; otherwise knowledge discovery is very difficult or can lead to misleading results. Practitioners who want to use SM mining to support their decision making processes need to select appropriate pre-processing steps, else the results might be inconclusive.

For this study, much more time was spent on data pre-processing than the effective data analysis. This is due to several facts that need to be considered if SM mining tasks using Twitter are conducted:

- Collecting tweets is time consuming due to the limitations of the Twitter API
- The collected data needs to be verified for its quality, which is a manual process and can only be partly automated
- To obtain a good dataset for training, many pre-processing and verification iterations were necessary
- ML and correlation techniques have reached a high level of maturity and many studies have been conducted to improve them

- ML techniques such as Logistic Regression are highly configurable and can be trained in a manageable amount of time
- There are many established data analysis frameworks that have been proven to deliver good results in many studies
- ML techniques are well documented in the literature
- Good ready-made training sets for specific DM tasks are rare and often it is laborious and time consuming and needs manual steps to create them
- Data pre-processing techniques differ greatly and are highly dependent on the task at hand, for example in the cases where financial data is analysed or text data is mined

Looking only at the performance values such as F-score and accuracy would not produce credible results and manual checks will be needed. The reason is that the training and test data are limited and random sampling was used to do manual post-training verification. A set of opinions, not found in the training and test corpus, was applied to the trained learner. Also, since many words have not been seen during training, the verification step has shown to give additional information about the performance of the trainer that the standard measures had not.

Some classifiers such as LR are highly configurable and to optimize them is very time consuming. Even for an experienced data scientist familiar with ML techniques the results are sometimes unpredictable and surprising. For instance, in LR, very counterintuitively character n-grams performed better than tokens.

Using ML methods as black boxes makes it difficult to understand why they fail in certain situations and how to fix them. Many classifiers are difficult to interpret since the inner state is complex. The implication is a loss of control over how the internals work, for

instance in deep learners, but also with shallow learners such as ensemble learners that are often difficult to interpret. Bayesian networks or decision trees are more transparent than a complex neural network or deep learner. Thus it might be difficult to determine how an AI system came to a certain conclusion.

The implications for organizations are that SM analysis should be applied only to specific tasks as part of an overall IT strategy, but should not be used as a single source of information.

5.5 Limitations

Arthur Lee Samuel coined the term “Machine Learning” in a paper in 1959 (Samuel 1959). Since then many ML techniques have been developed. Also data conditioning techniques have also been evolving alongside them. In this study some techniques were evaluated that had been widely used in previous studies in very different contexts (Xinyu, Youngwoon & Suk young 2015; Souza, TTP, Kolchyna, Treleaven & Aste 2015; Arias, Arratia & Xuriguera 2014; Bollen, Mao & Zeng 2010; Asur & Huberman 2010) and that showed to provide good results. Due to the large numbers of techniques, a decision was made, based on the studies analysed in the literature review, to use the techniques described in chapter 3. It was beyond the scope of this study to use all of the techniques. The same applied to the data pre-processing techniques. Some of the typical pre-processing steps such as data deduplication and basic subjectivity analysis were performed. Applying more data pre-processing steps such as lemmatisation or part-of-speech tagging might have improved the predictive accuracy. However due to the large

number of techniques available only some of the more popular ones were applied. Evaluating pre-processing techniques was beyond the scope of this study.

The subject of opinion mining using SM data is more complex than the literature suggests. Some studies mentioned in the literature review found a direct correlation between public mood states in SM and the Dow Jones Industrial Average (DJIA) (Bollen, Mao & Zeng 2010), or political opinions and election results (Tsakalidis et al. 2015). However, there are many more factors that influence the result. Tweets only mirror certain factors and are subject to the self-selection bias and possibly to the network effect.

Limitations were encountered at several stages of the research:

- Limitations in collecting the data
- Limitations inherent to opinion mining
- Limitations in the ML algorithms

There are several limitations in harvesting tweets. At the time of writing, the search API and Web based query interface did not render the same result set despite being based on relevance, with the streaming API only giving a random number of tweets. The Twitter documentation did not state which criteria are used to classify the relevance of the select tweets.

There are also limitations due to the complexity of the task of opinion mining, having to cater for word sense disambiguation, coreference resolution and negation handling. NLP remains a challenging task and this is a rapidly evolving field of study. More research in this area is needed to increase the accuracy.

5.6 Further research

New methods using deep learning have been shown to deliver very promising results. Particularly, two types of deep learners have performed exceptionally well on certain tasks: Recurrent Neural Networks (RNN) for NLP and Convolutional Neural Networks (CNN) for Multimedia Mining (Wlodarczak et al. 2015, p. 191). There is no agreed upon definition for deep learners, but they are typically artificial neural networks comprised of many layers and can proceed hierarchically from the input observations into more abstract levels of representation as they pass from one layer to the next. They take advantage of the hierarchical structures often found in nature. Natural language is composed of letters, letters form phonemes, phonemes form words, words form phrases, phrases form sentences etc. Contrary to many learners, RNN do not take a fixed size vector an input. This makes them ideal for NLP tasks. Natural language can be of varied sizes. RNN maintain an inner state through their recurrent layers. RNN are very suited for semantic analysis of texts of oral speeches.

One of the big advantages of deep learners is that they can automatically extract features. Deep learners are also more plausible biologically. Most human or animal learning is not supervised. We learn from experience, not from labelled data. On the down side, Deep Learning uses significantly more data pre-processing than shallow learners and they have a carnivorous appetite for data. Also, optimizing a Deep Learning Model can be very computationally intensive. Deep learners have been very successful because of advances in processor technologies such as GPUs (Graphics Processing Units) that generate a great deal of computing power. Using deep learners for opinion mining is a still fertile ground for future research.

Feature engineering usually consumes most of a DM task and more automation would be greatly beneficial. Feature engineering requires human judgement which makes it difficult to automate. Also, the heuristics are highly domain specific and the feature space is not always obvious. Since unsupervised learners do not require an input vector but operate on latent variables as causes for all observations, they can learn larger and more complex models. Latent variables cannot be directly observed but are inferred from observations. If the causal relation between the input and output is complex, supervised learning, in practice, cannot learn where deep hierarchies exist.

Deep Linguistic Analysis methods are considering the context contrary to the bag-of-words approach that many ML techniques use. It handles the structure of a language at the morphology, syntax and semantics level. However, linguistics does not look at the insights and ML techniques are better suited to extract knowledge.

Big Data analysis usually deals with large datasets. Since a smaller, Apple specific data set performed better from a large data set with opinions on other products and companies than Apple in this study, research on different types of datasets with different sizes could lead to better performing classifications. In this study, the datasets were pre-processed using basic subjectivity analysis and deduplication. A data set large enough might not need data purification according to Big Data principles where no high quality samples are needed. Analysing the predictive performance of different datasets in different sizes might lead to valuable results as to which lead to the best results.

Developing ML models is an iterative process that needs fine tuning. Automating tasks such as automated model selection will speed up the learning cycle. However, there is no universal effectiveness measure for all models. Some research has already provided results such as Auto-WEKA (Auto-WEKA 2016).

MLaaS (Machine Learning as a Service) solutions such as Google Cloud Prediction API (Google Cloud Platform 2016) will potentially multiply the ML applications. However, cloud based solutions often hide the complexity of the implementation, and, due to the black box approach, some control is lost, which makes it difficult to optimize the learners or interpret the results.

Properly establishing causality is a difficult endeavour and a subject for more research. The inference of causality based on empirical data requires:

- the putative causative factor to be the only variable factor in the experiment and its values must be completely under the control of the researcher
- the response to be evaluated in time immediately after the value of the putative causative factor is changed
- a reasonable model that explains the possible nature of the causative link between the factor on the response

The most effective way to identify causality is through a well-controlled experiment (de Siqueira Santos et al. 2013, p. 12). Future research on designing the experiment could thus lead to better correlation analysis and a better understanding of causality.

A new area of research is target-dependent SA, where the query represents the target of the query. Some studies have already developed classifiers using this paradigm. Their classifiers actually work in a target-independent way: all the features used in the classifiers are independent of the target, so the sentiment is decided no matter what the target is (Jiang et al. 2011, p. 151). Since the target, of for instance "iPhone", is clearly defined; basic subjectivity analysis in this study should suffice for target-dependant SA. However, in some cases there might be disambiguation. For instance, there is the popular "Game of Thrones" TV series, a novel by George R. R. Martin, and a game and

a comic with the same name. Tweets about "Game of Thrones" have to be first analysed in order to filter out the tweets about the desired target. Also, people may talk about multiple targets in a tweet. Target-dependant SA is thus an area for more research. Based on our manual evaluation of Twitter Sentiment output, about 40% of errors are because of having disambiguation (Jiang et al. 2011, p. 152). Taking into consideration the relations between tweets, such as tweets published by the same person, tweets replying to or replied by the given tweet, or retweets of a given tweet. These relations might provide information about what the given tweet expresses and could increase the accuracy of the classification. This allows for context-aware sentiment classification and might improve the performance of the learner.

Finally, text analysis can be expanded to virtually any domain that humans write about. For instance, text can be analysed for jurisdictional decisions or medical diagnosis based on anamnesis. However, there are ethical considerations if a computer decides on the guilt of a person or proposes a diagnosis. The ethical implications of ML are a potential new area of research given that their ramifications are not fully understood and are intensely debated.

5.7 Conclusions

Several ML schemes were trained and tested. While some delivered results with low accuracy, the classifiers that did perform well had similarly good predictive performance. For this study, during DM, more of the time had to be spent on data pre-processing, with less time left on the data analysis. Automatic feature extraction can thus reduce the DM effort considerably and more research in this area should be conducted. DM techniques,

such as deep learning, that have the capability to perform automatic feature extraction, have the potential to reduce the time spent on pre-processing tasks dramatically and make the entire DM cycle more efficient. However, while deep learning techniques have delivered surprisingly good results with tasks such as Multimedia Mining and Natural Language Processing, they were less effective on non-perceptual systems. They are also statistically less well understood in comparison to ML techniques and it is often difficult to determine how a result was obtained. Using deep learners, there is a loss of control over how the internals work, as is the case with other learners, like, for example, ensemble learners.

Determining causation in many correlation problems is difficult since often not all factors are known. This study attempted to correlate Twitter opinions with share price. Many different factors influence the development of financial markets and price movements, and much research has already been conducted to determine these factors. However, properly designing the experiment for causality is crucial to obtaining good results and further research is necessary to obtain and interpret results conclusively.

ML techniques are biologically not very plausible. To develop techniques that work more “brain-like”, a better understanding of how neural networks work in nature must be obtained to build systems that better represent learning in biological systems. Deep learners seem to be closer to nature, but they are still a long way from simulating how neurobiologists believe the human or animal brain works. Most human or animal learning is not supervised. We learn from experience, not from labelled data. Humans have their whole lifetime of experience which they can apply to solve problems. Computers do not have the quantity and quality of data that we accumulate over a lifetime. However, we have already seen tasks where humans, in the past, have outperformed computers (for

example, chess or Go, the strategy board game), but are now being beaten by machines in recent years. Deep learners have mastered tasks such as speech and object recognition with an ever-increasing accuracy. But the subconscious mind gives humans context and without knowing how the subconscious mind works, it is almost impossible to know how to implement it. The subconscious mind is far more complex than the conscious mind. Moravec's paradox states that high-level reasoning takes very little computational power contrary to low-level sensorimotor skills (Rotenberg 2013). Giving computers the capability of perception and mobility is currently an almost insurmountable task, and researchers still have to go a long way before computers will have the ability to achieve the skills of even a small child in this area.

6 Appendices

6.1 References

Abbruzzese, J 2013, ‘Marketers Learn to Play by Facebook's Changing Rules’, *Mashable*, viewed 2 January 2014, <http://mashable.com/2013/12/30/marketers-facebook-rules>.

Achrekar, H, Gandhe, A, Lazarus, R, Ssu-Hsin, Y & Benyuan, L 2011, 'Predicting Flu Trends using Twitter data', in Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on: proceedings of theComputer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on pp. 702-7.

Agrawal, D & Aggarwal, CC 2001, 'On the design and quantification of privacy preserving data mining algorithms', paper presented to Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Santa Barbara, California, USA.

Agrawal, R & Srikant, R 2000, 'Privacy-preserving data mining', *ACM Sigmod Record*, vol. 29, no. 2, pp. 439-50, viewed 29 September 2013, < http://www.ilunwen.com/translation/20120520/20120520121602kjzhdmpx.pdf>.

Akerkar, R 2005, *Introduction to Artificial Intelligence*, Prentice-Hall of India Pvt.Ltd, Delhi, India.

Akhtar, MM, Zamani, AS & El-Sayed, A 2012, 'Link Analysis using Data Mining System', *International Journal of Applied Research in Computer Science and Information Technology*, vol. 1, no. 2, pp. 38-49, viewed 1 May 2013, <http://www.setscholars.org/index.php/ijarcsit/article/view/118/50>.

Alimi, SR & Ofonyelu, CC 2013, 'Toda-Yamamoto causality test between money market interest rate and expected inflation: the Fisher hypothesis revisited', European Scientific Journal, ESJ, vol. 9, no. 7.

Al-Oufi, S, Kim, H-N & El Saddik, A 2012, 'A group trust metric for identifying people of trust in online social networks', *Expert Systems with Applications*, vol. 39, no. 18, pp. 13173-81.

Amazon Machine Learning, 2017, 'Amazon Machine Learning - Predictive Analytics with AWS', viewed 20 June 2017, <https://aws.amazon.com/machine-learning/>.

Apple Inc. 2016, 'Q3 2016 Unaudited Summary Data', viewed 11 September 2016, < https://www.apple.com/uk/pr/pdf/q3fy16datasum.pdf>.

Arias, M, Arratia, A & Xuriguera, R 2014, 'Forecasting with twitter data', *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 1, pp. 1-24.

Asay, M 2012, 'Open API lessons for LinkedIn and Facebook', *The Register*, UK, viewed 4 January 2014, <http://www.theregister.co.uk/2012/05/31/linkedin_closed_apis/>.

Ashley, RA & Tsang, KP 2014, 'Credible Granger-Causality Inference with Modest Sample Lengths: A Cross-Sample Validation Approach', Econometrics, vol. 2, p. 19.

Asur, S & Huberman, BA 2010, 'Predicting the Future with Social Media', in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 492-9.

Atzmueller, M 2012, 'Onto collective intelligence in social media: exemplary applications and perspectives', paper presented to Proceedings of the 3rd international workshop on Modeling social media, Milwaukee, Wisconsin, USA.

Auditore, PJ 2012, 'SOCIAL MEDIA AND BUSINESS INTELLIGENCE SURVEY RESULTS & ANALYSIS', *Unisphere Research*, viewed 30 April 2013, <http://www.gse.org/Portals/2/docs/GSE%20Docs/SocialMedia_BI_SurveyReport_Final.pdf>.

Austin, PC 2011, 'An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies', *National Center for Biotechnology Information*, viewed 9 May 2013, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/>.

Auto-WEKA 2016, 'Auto-WEKA', *The University of British Columbia*, viewed 20 November 2016, <http://www.cs.ubc.ca/labs/beta/Projects/autoweka/>.

Azevedo, A & Santos, MF 2012, 'Binding Data Mining to Final Business Users of Business Intelligence Systems', *The First International Conference on Intelligent Systems and Applications*, viewed 2 May 2013, <http://www.thinkmind.org/index.php?view=article&articleid=intelli_2012_1_20_80043>.

Back, M, Nestler, S, Egloff, B & Stopfer, J 2013, 'Facebook-Nutzer sind realistisch und ehrlich, Facebook user are realistic and honest', *Universität Münster*, viewed 9 Sept 2013, <http://www.uni-muenster.de/Rektorat/exec/upm.php?rubrik=Alle&neu=0&monat=201309&nummer=16972>.

Backstrom, L, Boldiy, P, Rosay, M, Ugander, J & Vigna, S 2012, 'Four Degrees of Separation', *Università degli Studi di Milano*, viewed 20 Aug 2013, <http://arxiv.org/abs/1111.4570>.

Barberá, P & Rivero, G 2013, 'Understanding the political representativeness of Twitter users', viewed 21 May 2014, <https://files.nyu.edu/pba220/public/barbera_rivero_2013.pdf>.

Bai, Z, Wong, W-K & Zhang, B 2010, 'Multivariate linear and nonlinear causality tests', *Mathematics and Computers in Simulation*, vol. 81, no. 1, pp. 5-17, <http://www.sciencedirect.com/science/article/pii/S0378475410001977>.

Berman, JJ 2013, *Principles of Big Data*, Elsevier Inc., Waltham, USA.

Bollen, J, Mao, H & Zeng, X-J 2010, 'Twitter mood predicts the stock market', *Journal of Computational Science*, vol. 2, p. 8.

Bouktif, S & Awad, MA 2013, 'Ant colony based approach to predict stock market movement from mood collected on Twitter', in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*: *proceedings of the Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* ACM, Niagara, Ontario, Canada, pp. 837-45.

Boyd, D & Crawford, K 2012, 'Critical Questions for Big Data', Information, Communication & Society, vol. 15, no. 5, pp. 662-79.

Baldwin, B & Dayanidhi, K 2014, *Natural Language Processing with Java and LingPipe Cookbook*, Packt Publishing, UK.

Bryman, A 1989, *Research Methods And Organization Studies*, Routledge, New York, USA.

Bryman, A & Bell, E 2007, *Business Research Methods*, 2nd edn, Oxford University Press, Oxford, USA.

Buhl, H, Röglinger, M, Moser, F & Heidemann, J 2013, 'Big Data', *WIRTSCHAFTSINFORMATIK*, vol. 55, no. 2, pp. 63-8.

Bulysheva, L & Bulyshev, A 2012, 'Segmentation modeling algorithm: a novel algorithm in data mining', *Information Technology and Management*, vol. 13, no. 4, pp. 263-71.

Burgess, J & Bruns, A 2012, 'Twitter Archives and the Challenges of "Big Social Data" for Media and Communication Research', *M/C Journal*, vol. 15, no. 5, viewed 27 September 2013, < http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/561>.

Burnap, P, Gibson, R, Sloan, L, Southern, R & Williams, M 2015, '140 Characters to Victory?: Using Twitter to Predict the UK 2015 General Election', Computers and Society.

Business 2 Community 2014, *Looking Back: Social Media Numbers in 2013*, Conduit Mobile, viewed 4 January 2014, <http://www.business2community.com/social-media/looking-back-social-media-numbers-2013-0728246#!rjI7z>.

Butler, D 2013, 'When Google got flu wrong', *Nature*, viewed 13 May 2013, <http://www.nature.com/news/when-google-got-flu-wrong-1.12413>.

Carpenter, B 2004, 'Phrasal queries with LingPipe and Lucene: ad hoc genomics text retrieval', in TREC: proceedings of theTREC pp. 1-10.

Carpenter, B 2007, 'LingPipe for 99.99% recall of gene mentions', in Proceedings of the Second BioCreative Challenge Evaluation Workshop: proceedings of theProceedings of the Second BioCreative Challenge Evaluation Workshop pp. 307-9.

Chan, JO 2013, 'An Architecture for Big Data Analytics', *Communications of the IIMA*, vol. 13, no. 2, pp. 1-13, <http://ezproxy.usq.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=iih&AN=95612792&site=ehost-live>.

Chau, M & Xu, J 2012, 'Business Intelligence in Blogs: Understanding consumer interactions and communities, MIS Quarterly, vol. 36, no. 4, pp. 1189-1216, viewed 14 May 2013, <http://www.fbe.hku.hk/~mchau/papers/BusinessIntelligenceInBlogs_MISQ.pdf>.

Chen, KL, Lee, H, Shing, C-C & Yang, J 2009, 'An Analysis of Algorithms Used by Business Intelligence Software ', *Proceedings of International Conference on Pacific Rim Management*, pp. 36-9.

Cheng-Zhang, P, Ze-Jun, J, Xiao-Bin, C & Zhi-Ke, Z 2012, 'Real-time analytics processing with MapReduce', in *Machine Learning and Cybernetics (ICMLC), 2012 International Conference on*: *proceedings of theMachine Learning and Cybernetics (ICMLC), 2012 International Conference on* pp. 1308-11.

Chui, M, Löffler, M & Roberts, R 2010, 'The Internet of Things', *McKinseyQuarterly*, viewed 23 April 2013, <http://www.mckinseyquarterly.com/The_Internet_of_Things_2538>.

Cioffi-Revilla, C 2010, 'Computational social science', *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 3, pp. 259-71.

Clemons, EK 2009, 'The complex problem of monetizing virtual electronic social networks', *Decision Support Systems*, vol. 48, no. 1, pp. 46-56, <http://www.sciencedirect.com/science/article/pii/S0167923609001298>.

Pramuk, J 2015, 'Apple earnings: $1.96 per share, vs $1.88 expected', *CNBC*, viewed 2 February 2016, < http://www.cnbc.com/2015/10/27/apple-q4-earnings-results.html>.

Collins, H 2010, *Creative Research: The Theory and Practice of Research for the Creative Industries*, 1st edn, AVA Publishing SA, Lausanne, Switzerland.

Corley, C, Cook, D, Mikler, A & Singh, K 2010, 'Text and Structural Data Mining of Influenza Mentions in Web and Social Media', *International journal of environmental research and public health*, vol. 7, no. 2, pp. 596-615.

Crampton, JW, Graham, M, Poorthuis, A, Shelton, T, Stephens, M, Wilson, MW & Zook, M 2013, 'Beyond the Geotag Deconstructing 'Big Data' and Leveraging the Potential of the Geoweb', *University of Kentucky*, viewed 26 September 2013, < http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2253918>.

Curd, M & Psillos, S 2013, *The Routledge Companion to Philosophy of Science*, 2nd edn, Routledge, New York, NY.

Dai, Y, Kakkonen, T & Sutinen, E 2011, 'SoMEST: a model for detecting competitive intelligence from social media', paper presented to Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, Tampere, Finland.

Damodaran, A 2013, 'The Data Page', *Stern School of Business New York*, viewed 24 December 2013, <http://pages.stern.nyu.edu/~%20adamodar/>.

Darwish, A & Lakhtaria, KI 2011, *The Impact of the New Web 2.0 Technologies in Communication, Development, and Revolutions of Societies*, vol. 2, 2011.

DataSift 2015, 'Human Data Intelligence', viewed 24 August 2015, <http://datasift.com>.

Denecke, K 2008, 'Using sentiwordnet for multilingual sentiment analysis', IEEE 24th International Conference on: proceedings of theData Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on IEEE, pp. 507-12.

de Siqueira Santos, S, Takahashi, DY, Nakata, A & Fujita, A 2013, 'A comparative study of statistical methods used to identify dependencies between gene expression signals', Briefings in Bioinformatics, pp. 1-13.

Dinu, B & Iovan, S 2014, 'Harnessing Big Data Volumes', *Fiability & Durability / Fiabilitate si Durabilitate*, no. 1, pp. 250-6, <http://ezproxy.usq.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=97069971&site=ehost-live>.

Du, Z 2013, 'Inconsistencies in big data', in Cognitive Informatics & Cognitive Computing (ICCI*CC), 2013 12th IEEE International Conference on: proceedings of theCognitive Informatics & Cognitive Computing (ICCI*CC), 2013 12th IEEE International Conference on pp. 61-7.

Ertekin, Ş, Rudin, C & Hirsh, H 2014, 'Approximating the crowd', *Data Mining and Knowledge Discovery*, vol. 28, no. 5-6, pp. 1189-221, <http://dx.doi.org/10.1007/s10618-014-0354-1>.

Evangelopoulos, N, Magro, MJ & Sidorova, A 2012, 'The dual micro/macro informing role of social network sites: can Twitter macro messages help predict stock prices?', Informing science, vol. 15, p. 247.

Facebook 2013, 'Key Facts', *Facebook*, viewed 28 October 2013, <http://newsroom.fb.com/Key-Facts>.

Fan, G-F, Qing, S, Wang, H, Hong, W-C & Li, H-J 2013, 'Support vector regression model based on empirical mode decomposition and auto regression for electric load forecasting', Energies, vol. 6, no. 4, pp. 1887-901.

Finlay, S., 2014, Predictive analytics, data mining and big data: myths, misconceptions and methods, Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, Houndmills, Basingstoke, Hampshire.

Fisher, C, Buglear, J, Lowry, D, Mutch, A & Tansley, C 2007, *Researching and Writing a Dissertation: A Guidebook for Business Students*, 2nd edn, Pearson Education Limited, England.

Garcia Martinez, M & Walton, B 2014, 'The wisdom of crowds: The potential of online communities as a tool for data analysis', *Technovation*, vol. 34, no. 4, pp. 203-14, <http://www.sciencedirect.com/science/article/pii/S0166497214000182>.

Gerber, MS 2014, 'Predicting crime using Twitter and kernel density estimation', *Decision Support Systems*, vol. 61, pp. 115-25, <http://www.sciencedirect.com/science/article/pii/S0167923614000268>.

Gilbert, N 2010, *Computational Social Science*, SAGE Publications Ltd, London, UK.

Ginsberg, J, Mohebbi, MH, Patel, RS, Brammer, L, Smolinski, MS & Brilliant, L 2009, 'Detecting influenza epidemics using search engine query data', *Nature*, vol. 457, no. 7232, pp. 1012-4, <http://dx.doi.org/10.1038/nature07634>.

Gnip 2015, 'The Source for Social Data', viewed 21 July 2015, <https://gnip.com/>.

Gokhale, C, Das, S, Doan, A, Naughton, JF, Rampalli, N, Shavlik, J & Zhu, X 2014, 'Corleone: hands-off crowdsourcing for entity matching', in Proceedings of the 2014 ACM SIGMOD international conference on Management of data: proceedings of the Proceedings of the 2014 ACM SIGMOD international conference on Management of data ACM, Snowbird, Utah, USA, pp. 601-12.

Google Cloud Platform 2016, ‘Prediction API - Pattern Matching in the Cloud’, *Google Cloud Platform*, viewed 20 November 2016, <https://cloud.google.com/prediction/>.

Goh, KY, Heng, CS & Lin, Z 2012, ‘Social Media Brand Community and Consumer Behavior: Quantifying the Relative Impact of User- and Marketer-Generated Content’, *School of Computing*, National University of Singapore, viewed 9 April 2013, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2048614>.

Gomez-Arias, JT & Genin, L 2009, 'BEYOND MONETIZATION: CREATING VALUE THROUGH ONLINE SOCIAL NETWORKS', *International Journal of Electronic Business Management*, vol. 7, no. 2, pp. 79-85, <http://ezproxy.usq.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=50738160&site=ehost-live>.

González-Recio, O, Rosa, GJM & Gianola, D 2014, 'Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits', *Livestock Science*, vol. 166, no. 0, pp. 217-31, <http://www.sciencedirect.com/science/article/pii/S1871141314003114>.

Governor, J, Hinchcliffe, D & Nickull, D 2009, *Web 2.0 Architectures: What Entrepreneurs and Information Architects Need to Know*, 1st edn, O'Reilly, Sebastopol, USA.

Graham, DM, Hale, SA & Stephens, M 2011, 'User-generated Content in Google', *Oxford University*, Oxford, UK, viewed 27 October 2013, < http://www.oii.ox.ac.uk/vis/?id=4e3c030d>.

Greenwald, G 2013, *NSA collecting phone records of millions of Verizon customers daily*, The Guardian, UK, viewed 4 January 2014, <http://www.theguardian.com/world/2013/dec/29/2013-eyewitness-acounts-edward-snowden>.

Groff, R 2004, *Critical Realism, Post-positivism and the Possibility of Knowledge*, Routledge, New York, USA.

Groppelli, AA & Nikbakht, E 2006, *Finance*, 5th edn, Barron's Educational Series, NY, USA.

Gundecha, P & Liu, H 2012, ‘Mining Social Media: A Brief Introduction’, *Arizona State University*, Tempe, Arizona, viewed 9 April 2013, <http://www.public.asu.edu/~pgundech/book_chapter/smm.pdf>.

Han, J & Kamber, M 2006, Data Mining: Concepts and Techniques, Morgen Kaufmann, Waltham, MA, USA.

Han, J, Kamber, M & Pei, J 2011, *Data Mining: Concepts and Techniques*, Morgen Kaufmann, Waltham, MA, USA.

Hangya, V & Farkas, R 2013, 'Target-oriented opinion mining from tweets', in Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on: proceedings of theCognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on pp. 251-4.

Harris, D 2013, 'DataSift raises $42M', *Gigaom*, viewed 27 December 2013, <http://gigaom.com/2013/12/03/datasift-raises-42m-maybe-theres-something-to-this-social-data-after-all/>.

Hiemstra, C & Jones, JD 1994, 'Testing for Linear and Nonlinear Granger Causality in the Stock Price- Volume Relation', *The Journal of Finance*, vol. 49, no. 5, pp. 1639-64, <http://www.jstor.org/stable/2329266>.

Higgins, T 2015, 'Apple IPhone Shipments, Revenue Forecast Miss Estimates', *Bloomberg*, viewed 30 March 2016, <http://www.bloomberg.com/news/articles/2015-07-21/apple-iphone-shipments-revenue-forecast-miss-estimates>.

Hong Keel, S, Dennis, AR & Yuan, LI 2014, 'Trading on Twitter: The Financial Information Content of Emotion in Social Media', in System Sciences (HICSS), 2014 47th Hawaii International Conference on: proceedings of theSystem Sciences (HICSS), 2014 47th Hawaii International Conference on pp. 806-15.

Huang, G, Huang, G-B, Song, S & You, K 2015, 'Trends in extreme learning machines: A review', Neural Networks, vol. 61, pp. 32-48, <http://www.sciencedirect.com/science/article/pii/S0893608014002214>.

Huang, S, Peng, W, Li, J & Lee, D 2013, 'Sentiment and topic analysis on social media: a multi-task multi-label classification approach', paper presented to Proceedings of the 5th Annual ACM Web Science Conference, Paris, France, DOI 10.1145/2464464.2464512.

Hochman, N & Manovich, L 2013, 'Zooming into an Instagram City: Reading the local through social media', first monday, peer-reviewed journal on the internet, vol. 18, no. 7, viewed 30 Aug 2013, < http://journals.uic.edu/ojs/index.php/fm/article/view/4711/3698>.

Hopkins, D 2005, 'Heckman Selection Models', *Washington University*, St. Louis, viewed 13 May 2013, <http://rtm.wustl.edu/GMMC/heckman.pdf>.

Ishijima, H, Kazumi, T & Maeda, A 2015, 'Sentiment analysis for the Japanese stock market', Global Business and Economics Review, vol. 17, no. 3.

Jackson, S 2012, *Research Methods and Statistics: A Critical Thinking Approach*, 2 edn, Wadsworth Cengage Learning, Belmont, CA, USA.

Jadav, JJ & Panchal, M 2012, 'Association Rule Mining Method On OLAP Cube', *International Journal of Engineering Research and Applications*, vol. 2, no. 2, viewed 6 May 2013, <http://www.ijera.com/papers/Vol2_issue2/GL2211471151.pdf>.

Jagadish, HV, Gehrke, J, Labrinidis, A, Papakonstantinou, Y, Patel, JM, Ramakrishnan, R & Shahabi, C 2014, 'Big data and its technical challenges', Commun. ACM, vol. 57, no. 7, pp. 86-94.

Jiang, L, Yu, M, Zhou, M, Liu, X & Zhao, T 2011, 'Target-dependent Twitter sentiment classification', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* Association for Computational Linguistics, Portland, Oregon, pp. 151-60.

Jones, I & Huan, L 2013, 'Mining Social Media: Challenges and Opportunities', in *Social Intelligence and Technology (SOCIETY), 2013 International Conference on*, pp. 90-9.

Kabacoff, RI 2011, R in Action, Manning Publications Co., Shelter Island, NY.

Kalampokis, E, Tambouris, E & Tarabanis, K 2013, 'Understanding the predictive power of social media', *Internet Research*, vol. 23, no. 5, pp. 544-59.

Kao, A, Ferng, W, Poteet, S, Quach, L & Tjoelker, R 2013, 'TALISON - Tensor analysis of social media data', in *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, pp. 137-42.

Klein, D, Tran-Gia, P & Hartmann, M 2013, 'Big Data', *Informatik-Spektrum*, vol. 36, no. 3, p. 319.

Konchady, M 2008, Building Search Applications: Lucene, LingPipe, and Gate, Mustru Publishing, Oakton VA.

Kosinski, M, Stillwell, D & Graepel, T 2013, 'Private traits and attributes are predictable from digital records of human behavior', *PNAS*, viewed 14 May 2013, < http://www.pnas.org/content/early/2013/03/06/1218772110.full.pdf+html>.

Kothari, CR 2004, *Research Methodology: Methods And Techniques*, 2nd edn, New Age International, New Delhi, India.

Kubr, M & Prokopenko, J 1989, 'Diagnosing Management Training and Development Needs: Concepts and Techniques', *Management Development Series*, no. 27, International Labour Office, Geneva, Switzerland.

Kumar, P, Nitin, Chauhan, DS & Sehgal, VK 2012, 'Selection of evolutionary approach based hybrid data mining algorithms for decision support systems and business intelligence', paper presented to Proceedings of the International Conference on

Advances in Computing, Communications and Informatics, Chennai, India, DOI 10.1145/2345396.2345563.

Kumar, P, Kumar Sehgal, N, Kumar Sehgal, V & Singh Chauhan, D 2012, 'A Benchmark to Select Data Mining Based Classification Algorithms for Business Intelligence and Decision Support Systems', *International Journal of Data Mining & Knowledge Management Process*, vol. 2, no. 5, pp. 25-42.

Laudon, KC & Laudon JP 2009, *Management Information Systems,* 11th edn, Pearson/Prentice Hall, Upper Saddle River.

Lazer, D, Kennedy, R, King, G & Vespignani, A 2014, 'The Parable of Google Flu: Traps in Big Data Analysis', *Science*, vol. 343, no. 14 March, pp. 1203-5.

Lazer, D, Pentland, A, Lada Adamic, Aral, S, Barabási, A-L, Brewer, D, Christakis, N, Contractor, N, Fowler, J, Gutmann, M, Jebara, T, King, G, Macy, M, Roy, D & Alstyne, MV 2009, 'Computational Social Science', *Science*, vol. 323, no. 5915, pp. 721-723.

LeCun, Y, Bengio, Y & Hinton, G 2015, 'Deep learning', Nature, vol. 521, no. 7553, pp. 436-44, <http://dx.doi.org/10.1038/nature14539>.

Lee, D, Ojo, A & Waqar, M 2013, 'Utilising Linked Social Media Data for Tracking Public Policy and Services', *Digital Enterprise Research Institute (DERI)*, NUI Galway, Ireland, viewed 12 May 2013, <http://www.w3.org/2013/04/odw/odw13_submission_39.pdf>.

Leetaru, K 2011, *Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space*, 2011.

Letichevsky, AA, Lyaletski, AV & Morokhovets, MK 2013, 'Glushkov's evidence algorithm', *Cybernetics and Systems Analysis*, vol. 49, no. 4, pp. 489-500.

Li, X, Xie, H, Chen, L, Wang, J & Deng, X 2014, 'News impact on stock price return via sentiment analysis', *Knowledge-Based Systems*.

Liang, P-W & Dai, B-R 2013, 'Opinion Mining on Social Media Data', in proceedings of the IEEE 14th International Conference on Mobile Data Management Italy, pp. 91-6, viewed <http://doi.ieeecomputersociety.org/10.1109/MDM.2013.73>.

Lim, E-P, Chen, H & Chen, G 2013, 'Business Intelligence and Analytics: Research Directions', *ACM Trans. Manage. Inf. Syst.*, vol. 3, no. 4, pp. 1-10.

Lin, N., Burt, R.S. & Cook, K.S. 2001, Social capital: theory and research, Aldine De Gruyter, New York.

LingPipe 2015, *LingPipe Home*, viewed 24 August 2015, <http://alias-i.com/lingpipe/>.

LingPipe API 2016, *LingPipe API*, viewed 29 March 2016, <http://alias-i.com/lingpipe/docs/api/index.html>.

Liu, B 2012, 'Opinion Mining', *University of Illinois at Chicago*, viewed 7 May 2013, <http://www.cs.uic.edu/~liub/FBS/opinion-mining.pdf>.

Liu, B 2012, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, USA.

Liu, T, Ding, X, Chen, Y, Chen, H & Guo, M 2014, 'Predicting movie Box-office revenues by exploiting large-scale social media content', *Multimedia Tools and Applications*, pp. 1-20, <http://dx.doi.org/10.1007/s11042-014-2270-1>.

Lloret, E, Balahur, A, Gómez, J, Montoyo, A & Palomar, M 2012, 'Towards a unified framework for opinion retrieval, mining and summarization', *Journal of Intelligent Information Systems*, vol. 39, no. 3, pp. 711-47, <http://dx.doi.org/10.1007/s10844-012-0209-4>.

Lubbadeh, J 2014, 'Seismograf der Welt', *Technology Review*, vol. 2., viewed 3 March 2014, <http://www.heise.de/tr/artikel/Seismograf-der-Welt-2096990.html>.

MALLET, 2017, 'MALLET homepage', viewed 20 June 2017, <http://mallet.cs.umass.edu/>.

Manyika, J, Chui, M, Brown, B, Bughin, J, Dobbs, R, Roxburgh, C & Byers, AH 2011, Big Data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute.

Marsland, S 2009, *Machine Learning: An Algorithmic Perspective*, Chapman & Hall, Boca Raton, FL, USA.

Mayer, A 2009, 'Online social networks in economics', *Decision Support Systems*, vol. 47, no. 3, pp. 169-184, viewed 22 September 2013, < http://sistemas-humano-computacionais.wdfiles.com/local--files/capitulo%3Aredes-sociais/amayer.pdf>.

Mayer-Schonberger, V & Cukier, K 2013, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt Publishing Company, New York, USA.

McKelvey, K, Rudnick, A, Conover, MD & Menczer, F 2012, 'Visualizing Communication on Social Media, Making Big Data Accessible', *Indiana University School of Informatics and Computing*, viewed 29 September 2013, <http://arxiv.org/pdf/1202.1367v1.pdf>.

McLaney, E & Atrill, P 2008, *Accounting, An Introduction*, 4th edn, Pearson Education Limited, Essex, UK.

Memon, N, Xu, JJ, Hicks, DL & Chen, H 2010, 'Data Mining for Social Network Data', *Annals of Information Systems*, Springer, vol. 12, pp. 1-9, viewed 18 May 2013, <http://www.springer.com/business+%26+management/business+information+systems/book/978-1-4419-6286-7>.

Menasalvas, E & Wasilewska, A 2006, 'Data Mining as Generalization: A Formal Model', in T Young Lin, et al. (eds), Foundations and Novel Approaches in Data Mining, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 99-126.

Meredith, R & O'Donnell, P 2012,' A Framework for Understanding the Role of Social Media in Business Intelligence Systems', *Monash University*, Melbourne, viewed 14 May 2013, <http://www.tandfonline.com/doi/abs/10.3166/jds.20.263-282#.UZH7eLX-GSo>.

Moore, H & Roberts, D 2013, 'AP Twitter hack causes panic on Wall Street and sends Dow plunging', *The Guardian*, viewed 24 April 2013, <http://www.guardian.co.uk/business/2013/apr/23/ap-tweet-hack-wall-street-freefall>.

Morgan, SL & Winship, C 2007, 'Counterfactuals and Causal Inference: Methods and Principles for Social Research', Cambridge University Press, New York, USA.

Mozenda Blog, 2017. 'Web Scraping Software', viewed 20 June 2017, <http://www.mozenda.com/>.

Nanli, Z, Yibo, W, Cheng, C, Wei, X, Yongping, Z, Ping, Z & Awan, MSK 2014, 'Regression-Based Microblogging Influence Detection Framework for Stock Market', Journal of Networks, vol. 9, no. 8, pp. 2129-36, <http://ezproxy.usq.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=iih&AN=97638899&site=ehost-live>.

Nasdaq, 2016, viewed 23 March 2016, <http://www.nasdaq.com/>.

Natural Language Toolkit, 2017, 'Natural Language Toolkit — NLTK 3.2.4 documentation', viewed 20 June 2017, <http://www.nltk.org/>.

Neri, F, Aliprandi, C, Capeci, F, Cuadros, M & By, T 2012, 'Sentiment Analysis on Social Media', in *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pp. 919-26.

Nguyen, NP, Yan, G & Thai, MT 2013, 'Analysis of misinformation containment in online social networks', *Computer Networks*, vol. 57, no. 10, pp. 2133-46, <http://www.sciencedirect.com/science/article/pii/S1389128613001126>.

Nijholt, A, Stock, O & Nishida, T 2009, 'Social intelligence design in ambient intelligence', *AI & SOCIETY*, vol. 24, no. 1, pp. 1-3.

NVivo 11 for Windows Help, 2016, 'How auto coding sentiment works', viewed 27 March 2016, <http://help-nv11.qsrinternational.com/desktop/concepts/How_auto_coding_sentiment_works.htm>.

Nohuddin, PNE, Coenen, F, Christley, R, Setzkorn, C, Patel, Y & Williams, S 2012, 'Finding "interesting" trends in social networks using frequent pattern mining and self organizing maps', Knowledge-Based Systems, vol. 29, pp. 104-13, viewed 14 Aug 2013, < http://www.sciencedirect.com/science/article/pii/S0950705111001420>.

Oboler, A, Welsh, K & Cruz, L 2012, The danger of Big Data: Social media as computational social science, *First Monday*, vol. 17, no. 7, viewed 15 Sep 2013, <http://journals.uic.edu/ojs/index.php/fm/article/view/3993/3269>.

O'Callaghan, D, Greene, D, Conway, M, Carthy, J & Cunningham 2013, 'Uncovering the Wider Structure of Extreme Right Communities Spanning Popular Online Networks', *Cornell University*, viewed 16 May 2013, <http://arxiv.org/pdf/1302.1726v1.pdf>.

Oinas-Kukkonen, H, Lyytinen, K & Yoo, Y 2010, 'Social Networks and Information Systems: Ongoing and Future Research Streams', *Journal of the Association for Information Systems*, vol. 11, no. 2, pp. 61-68, viewed 14 May 2013, <http://dblab.mgt.ncu.edu.tw/%E6%95%99%E6%9D%90/992-seminar/20.pdf>.

Olive, J, Christianson, C & McCary, J 2011, *Handbook of Natural Language Processing and Machine Translation*, Springer, New York, USA.

Oliveira, N, Cortez, P & Areal, N 2013, 'Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter', in *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*: *proceedings of the Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics* ACM, Madrid, Spain, pp. 1-8.

Olson, DL & Delen, D 2008, *Advanced Data Mining Techniques*, Springer-Verlag, Berlin, Germany.

Ostrowski, DA 2011, 'Predictive Semantic Social Media Analysis', in *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pp. 283-90.

Package 'lmtest', 2017, 'Testing Linear Regression Models', viewed 26 June 2017, < https://cran.r-project.org/web/packages/lmtest/lmtest.pdf>.

Paler-Calmorin, L & Calmorin, MA 1997, *Statistics in Education and the Sciences*, RBSI, Manila, Philippines.

Paltoglou, G & Thelwall, M 2012, 'Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media', *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 1-19.

Pang, B & Lee, L 2004, 'A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts', in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics: proceedings of the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics Association for Computational Linguistics, Barcelona, Spain, p. 271.

Pang, B & Lee, L 2008, 'Opinion Mining and Sentiment Analysis', *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1-2, pp. 1-137, viewed 28 September 2013, < http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf>.

Pardeep, K, Kumar, N, Sehgal, VK & Chauhan, DS 2012, 'A Benchmark to Select Data Mining Based Classification Algorithms for Business Intelligence and Decision Support Systems', *International Journal of Data Mining & Knowledge Management Process*, vol. 2, no. 5, pp. 25-42, <http://arxiv.org/ftp/arxiv/papers/1210/1210.3139.pdf>.

Peng, F & Schuurmans, D 2003, 'Combining naive Bayes and n-gram language models for text classification', in Proceedings of the 25th European conference on IR research: proceedings of the Proceedings of the 25th European conference on IR research Springer-Verlag, Pisa, Italy, pp. 335-50.

Peng, F, Schuurmans, D & Wang, S 2004, 'Augmenting Naive Bayes Classifiers with Statistical Language Models', Information Retrieval, vol. 7, no. 3, pp. 317-45, <http://dx.doi.org/10.1023/B:INRT.0000011209.19643.e2>.

Perera, C, Zaslavsky, A, Christen, P & Georgakopoulos, D 2014, 'Context Aware Computing for The Internet of Things: A Survey', Communications Surveys & Tutorials, IEEE, vol. 16, no. 1, pp. 414-54.

Perry, C 2013, *Efficient and Effective Research*, Work-applied learning series, AIB Publications, Adelaide, Australia.

Perry, C, Riege, A & Brown, L 1998, 'Realism rules ok: Scientific paradigms in marketing research about networks', *anzmac*, viewed 25 April 2013, <http://www.anzmac.org/conference/1998/Cd_rom/Perry73.pdf>.

Petz, G, Karpowicz, M, Fürschuß, H, Auinger, A, Stříteský, V & Holzinger, A 2014, 'Computational approaches for mining user's opinions on the Web 2.0', *Information Processing & Management*, vol. 50, no. 6, pp. 899-908, <http://www.sciencedirect.com/science/article/pii/S030645731400065X>.
Poole, DL & Mackworth, AK 2010, *Artificial Intelligence: Foundations of Computational Agents*, Cambridge University Press, New York, USA.

Porshnev, A, Redkin, I & Shevchenko, A 2013, 'Machine Learning in Prediction of Stock Market Indicators Based on Historical Data and Data from Twitter Sentiment Analysis', in

Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on: proceedings of theData Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on pp. 440-4.

Prem, J & Tate, S 2013, *Big Data Networked Storage Solution for Hadoop*, 1st ed. edn, Redpaper, IBM International Technical Support Organization, United States.

Preotiuc-Pietro, D & Cohn, T 2013, 'Mining User Behaviours: A Study of Check-in Patterns in Location Based Social Networks', *University of Sheffield*, viewed 3 May 2013, <https://staffwww.dcs.shef.ac.uk/people/D.Preotiuc/foursq13websci.pdf>.

Probst, F, Grosswiele, L & Pfleger, R 2013, 'Who will lead and who will follow: Identifying Influential Users in Online Social Networks', *Business & Information Systems Engineering*, vol. 5, no. 3, pp. 179-93, <http://dx.doi.org/10.1007/s12599-013-0263-7>.

'Propensity Score Matching in Observational Studies', *University of Manitoba*, viewed 13 May 2013,
<https://umanitoba.ca/faculties/medicine/units/community_health_sciences/departmental_units/mchp/protocol/media/propensity_score_matching.pdf>.

Puschmann, C & Burgess, J 2013, 'The Politics of Twitter Data', HIIG Discussion Paper Series, no. 2013-01 p. 15, viewed 5 Aug 2013,
< http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2206225>.

Qian, H 2009, On Data-driven Chi Square Statistics, ProQuest, MI, USA.

Qualman, E 2013, *Socialnomics: How Social Media Transforms the Way We Live and Do Business*, John Wiley & Sons, Inc., 2nd edn, New Jersy, USA.

RapidMiner, 2017, 'Data Science Platform', viewed 20 June 2017,
<https://rapidminer.com/>.

Ratnaparkhi, A 1997, A Simple Introduction to Maximum Entropy Models for Natural Language Processing, University of Pennsylvania, Philadelphia,
<http://repository.upenn.edu/cgi/viewcontent.cgi?article=1083&context=ircs_reports>.

Rotenberg, VS 2013, 'MORAVEC'S PARADOX: CONSIDERATION IN THE CONTEXT OF TWO BRAIN HEMISPHERE FUNCTIONS', *Activitas Nervosa Superior*, vol. 55, no. 3, p. 108.

Rubin, DB 2001, 'Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation', *Kluwer Academic Publishers*, viewed 13 May 2013,
<http://archlab.gmu.edu/people/jthompsz/rubin.pdf>.

Rui, H & Whinston, A 2011, 'Designing a Social-Broadcasting-Based Business Intelligence System', *The University of Texas at Austin*, viewed 14 May 2013,

<http://cism.mccombs.utexas.edu/attachments/154_Designing%20a%20Social-Broadcasting-Based%20Business%20Intelligence%20System.pdf>.

Rusli, EM 2013, 'Facebook Woos TV Networks With Data', *Digits*, viewed 15 February 2014, <http://blogs.wsj.com/digits/2013/09/29/facebook-woos-tv-networks-with-more-data/>.

Samuel, AL 1959, 'Some studies in machine learning using the game of checkers', IBM J. Res. Dev., vol. 3, no. 3, pp. 210-29.

Samsung Electronics Annual Report, 2012, Samsung, viewed 25 December 2013, <http://www.samsung.com/us/aboutsamsung/investor_relations/financial_information/downloads/2013/SECAR2012_Eng_Final.pdf>.

Saunders, M, Lewis, P & Thornhill, A 2009, *Research methods for business students*, 5th edn, Pearson Education Limited, Essex, UK.

Schmidt, MD & Lipson, H 2007, 'Learning noise', in Proceedings of the 9th annual conference on Genetic and evolutionary computation: proceedings of the Proceedings of the 9th annual conference on Genetic and evolutionary computation ACM, London, England, pp. 1680-5.

Schoen, H, Gayo-Avello, D, Metaxas, PT, Mustafaraj, E, Strohmaier, M & Gloor, P 2013, 'The power of prediction with social media', *Internet Research*, vol. 23, no. 5, pp. 528 - 43.

Schreck, T & Keim, D 2013, 'Visual Analysis of Social Media Data', *Computer*, vol. 46, no. 5, pp. 68-75.

Sentiment140 2016, *A Twitter Sentiment Analysis Tool*, viewed 26. April 2016, <http://help.sentiment140.com/home>.

Shen, D, Sun, J-T, Yang, Q & Chen, Z 2006, 'Text classification improved through multigram models', in Proceedings of the 15th ACM international conference on Information and knowledge management, Arlington, Virginia, USA, pp. 672-81.

Shulong, T, Yang, L, Huan, S, Ziyu, G, Xifeng, Y, Jiajun, B, Chun, C & Xiaofei, H 2014, 'Interpreting the Public Sentiment Variations on Twitter', *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 5, pp. 1158-70.

Shvachko, K, Hairong, K, Radia, S & Chansler, R 2010, 'The Hadoop Distributed File System', in *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*: *proceedings of theMass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on* pp. 1-10.

Siegel, E 2013, *Predictive analytics*, John Wiley & Sons, Hoboken, New Jersey, USA.

Siganos, A, Vagenas-Nanos, E & Verwijmeren, P 2014, 'Facebook's daily sentiment and international stock markets', *Journal of Economic Behavior & Organization*, <http://www.sciencedirect.com/science/article/pii/S0167268114001735>.

Smith, MS, Ventura, AD, Dewey, DP, Knutson, CD & Embley, DW 2011, 'A Computational Framework for Social Capital in Online Communities', *Brigham Young University*, viewed 28 July 2013, <http://posts.smithworx.com/publications/d.pdf>.

Solakidis, GS, Vavliakis, KN & Mitkas, PA 2014, 'Multilingual Sentiment Analysis Using Emoticons and Keywords', in Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on: proceedings of theWeb Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on pp. 102-9.

Souza, TTP, Kolchyna, O, Treleaven, PC & Aste, T 2015, 'Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry', Computers and Society, <http://arxiv.org/pdf/1507.00784v3.pdf>.

Spredfast, 2017, 'Social Media Experience Management Software Platform', viewed 20 June 2016, <https://www.spredfast.com/>.

Statista 2016, 'Twitter: number of active users 2010-2016', *Statista*, viewed 20 November 2016, <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>.

Tanaka, N, Giovannini, E, Witherell, W & Metzger, JM 2005, *Measuring Globalisation: OECD Handbook on Economic Globalisation Indicators*, OECD Publishing, Paris, France.

Tang, J, Chang, Y & Liu, H 2014, 'Mining social media with social theories: a survey', *SIGKDD Explor. Newsl.*, vol. 15, no. 2, pp. 20-9.

Ting, I-H, Wu, H-J & Ho, T-H 2010, 'Mining and Analyzing Social Networks', Studies in Computational Intelligence, vol. 288.

Topsy.com 2015, 'Twitter Search, Monitoring, & Analytics', http://topsy.com/viewed *21* July 2015, <http://topsy.com>.

Tretyakov, K 2004, 'Machine Learning Techniques in Spam Filtering', *Data Mining Problem-oriented Seminar*, p. 60-79, Estonia.

Trif, S 2011, 'Using Genetic Algorithms in Secured Business Intelligence Mobile Applications', *Informatica economica*, vol. 15, no. 1, pp. 69-79.

Trusov, M, Bucklin, RE & Pauwels, K 2009, Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site, *Journal of Marketing*, vol. 73, pp. 90–102, viewed 14 May 2013, <http://www.journals.marketingpower.com/doi/pdf/10.1509/jmkg.73.5.90>.

Tsakalidis, A, Papadopoulos, S, Cristea, AI & Kompatsiaris, Y 2015, 'Predicting Elections for Multiple Countries Using Twitter and Polls', *Intelligent Systems, IEEE*, vol. 30, no. 2, pp. 10-7.

Tsvetovat, M, Kazil, J & Kouznetsov, A, 2013, 'Implicit sentiment mining', *Institute for Operations Research and the Management Sciences*, vol. 39, no. 6, viewed 4 May 2013, <https://www.informs.org/ORMS-Today/Public-Articles/December-Volume-39-Number-6/Implicit-sentiment-mining>.

Tucker, P 2013, 'Has Big Data Made Anonymity Impossible?', *MIT Technology Review*, vol. 116, no. 4.

Tumasjan, A, Welpe, IM, Sandner, PG, Tumasjan, A & Sprenger, TO 2011, 'Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape', *Social science computer review*, vol. 29, no. 4, pp. 402-18.

Twinkle, A & Paul, S 2014, 'Adressing Big Data with Hadoop', *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 2, pp. 459–62.

Twitter Blogs 2015, *#numbers*, viewed 24 August 2015, <https://blog.twitter.com/2011/numbers>.

Twitter Developers 2017, *REST API*, viewed 26 Juni 2017, <https://dev.twitter.com/rest/public>.

Twitter Developers 2015, *The Search API*, viewed 24 August 2015, <https://dev.twitter.com/rest/public/search>.

Twitter4j 2017, ‘Introduction’, viewed 29 June 2017, < http://twitter4j.org/en/>.

Van Horne, JC & Wachowicz, JM 2008, *Fundamentals of financial Management*, 13th edn, Prentice Hall, Essex, UK.

Vatsalan, D, Christen, P & Verykios, VS 2013, 'A taxonomy of privacy-preserving record linkage techniques', Information Systems, vol. 38, no. 6, pp. 946-69, <http://www.sciencedirect.com/science/article/pii/S0306437912001470>.

Vercellis, C 2010, *Business Intelligence: Data Mining and Optimization for Decision Making*, John Wiley & Sons, West Essex, UK.

Vesdapunt, N, Bellare, K & Dalvi, N 2014, 'Crowdsourcing algorithms for entity resolution', Proc. VLDB Endow., vol. 7, no. 12, pp. 1071-82.

Vogt, WP, Gardner, DC & Haeffele, LM 2012, *When to Use What Research Design*, The Guilford Press, NY, USA.

Vuori, V & Okkonen, J 2012, 'Refining information and knowledge by social media applications', *Tampere University of Technology*, viewed 14 May 2013, <https://www.emeraldinsight.com/journals.htm?articleid=17015048>.

Wang, F-Y, Carley, KM, Zeng, D & Mao, W 2007, 'Social Computing: From Social Informatics to Social Intelligence', *Intelligent Systems, IEEE*, vol. 22, no. 2, pp. 79-83.

Wei, W, Mao, Y & Wang, B 2015, 'Twitter volume spikes and stock options pricing', *Computer Communications*, Elsevier, <http://www.sciencedirect.com/science/article/pii/S0140366415002340>.

Weka 3 2015, *Data Mining with Open Source Machine Learning Software in Java,* viewed 24 August 2015, <http://www.cs.waikato.ac.nz/ml/weka/>.

Welcome to Twitter 2015, *Login or Sign up*, viewed 24 August 2015, <https://twitter.com/>.

Welsh, E 2002, Dealing with Data: Using NVivo in the Qualitative Data Analysis Process, vol. 3.

Willett, W, Heer, J & Agrawala, M 2012, 'Strategies for crowdsourcing social data analysis', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: *proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* ACM, Austin, Texas, USA, pp. 227-36.

Witten, IH, Frank, E & Hall, MA 2011, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn, Morgan Kaufmann Publishers, Burlington, USA.

Wlodarczak, P, Qian, S, Ally, M & Soar, J 2015, 'Social genome mining for crisis prediction', in *21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining: proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, P Christen (ed.), Sydney.

Wlodarczak, P, Soar, J & Ally, M 2015, 'Genome Mining Using Machine Learning Techniques', in A Geissbühler, et al. (eds), *Inclusive Smart Cities and e-Health*, Springer International Publishing, vol. 9102, ch 39, pp. 379-84.

Wlodarczak, P, Soar, J & Ally, M 2015, 'Multimedia data mining using deep learning', in *Digital Information Processing and Communications (ICDIPC)*, 2015 Fifth International Conference on, IEEE Xplore, Sierre, pp. 190-6.

Wlodarczak, P, Soar, J & Ally, M 2015, 'Reality Mining in eHealth', in X Yin, et al. (eds), *Health Information Science*, Springer International Publishing, vol. 9085, ch 1, pp. 1-6.

Wlodarczak, P, Soar, J & Ally, M 2015, 'What the future holds for Social Media data analysis', *World Academy of Science, Engineering and Technology*, vol. 9, no. 1, p. 545.

Wong, FMF, Sen, S & Chiang, M 2012, 'Why Watching Movie Tweets Won't Tell the Whole Story?', *Cornell University*, viewed 14 May 2013, <http://arxiv.org/pdf/1203.4642v1.pdf>.

Wu, J & Coggeshall, S 2012, *Foundations of Predictive Analytics*, CRC Press, Boca Raton, FL, USA.

Wright, A 2014, 'Big Data Meets Big Science', *Communications of the ACM*, vol. 57, no. 7, pp. 13-5, <http://ezproxy.usq.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=96866646&site=ehost-live>.

Wu, X, Kumar, V, Ross Quinlan, J, Ghosh, J, Yang, Q, Motoda, H, McLachlan, GJ, Ng, A, Liu, B, Yu, PS, Zhou, Z-H, Steinbach, M, Hand, DJ & Steinberg, D 2007, 'Top 10 algorithms in data mining', Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37.

Xinyu, C, Youngwoon, C & Suk young, J 2015, 'Crime prediction using Twitter sentiment and weather', in *Systems and Information Engineering Design Symposium (SIEDS), 2015: proceedings of theSystems and Information Engineering Design Symposium (SIEDS), 2015* pp. 63-8.

Xu, K, Li, J & Song, Y 2012, 'Identifying valuable customers on social networking sites for profit maximization', *Elsevier*, viewed 14 May 2013, <https://www.sciencedirect.com/science/article/pii/S0957417412008147>.

Xu, Z, Tresp, V, Rettinger, A & Kersting, K 2010, 'Social Network Mining with Nonparametric Relational Models', in L Giles, M Smith, J Yen & H Zhang (eds), *Advances in Social Network Mining and Analysis*, Springer Berlin Heidelberg, vol. 5498, pp. 77-96, DOI 10.1007/978-3-642-14929-0_5, <http://dx.doi.org/10.1007/978-3-642-14929-0_5>.

Yom-Tov, E, Borsa, D, Cox, IJ & McKendry, RA 2014, 'Detecting Disease Outbreaks in Mass Gatherings Using Internet Data', *Journal of Medical Internet Research*, vol. 16, no. 6, p. e154, <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4090384/>.

Young, SD 2014, 'Behavioral insights on Big Data: using social media for predicting biomedical outcomes', *Trends in Microbiology*, vol. 22, no. 11, pp. 601-2, <http://www.sciencedirect.com/science/article/pii/S0966842X14001607>.

Zafarani, R, Abbasi, MA & Liu, H 2014, *Social Media Mining*, Cambridge University Press, New York.

Zeng, D, Chen, H, Lusch, R & Li, S-H 2010, 'Social Media Analytics and Intelligence', *Intelligent Systems, IEEE*, vol. 25, no. 6, pp. 13-6.

Zeng, L, Li, L & Duan, L 2012, 'Business intelligence in enterprise computing environment', *Information Technology and Management*, vol. 13, no. 4, pp. 297-310.

Zhang, K 2013, 'Big social media data mining for marketing intelligence', 3563913 thesis, Northwestern University, via ProQuest Dissertations & Theses A&I, viewed 3 November 2013, <http://search.proquest.com/docview/1400499921?accountid=14647>.

Zheng, C & Xiaoqing, D 2013, 'Study of Stock Prediction Based on Social Network', in Social Computing (SocialCom), 2013 International Conference on: proceedings of theSocial Computing (SocialCom), 2013 International Conference on pp. 913-6.

Zhong, N, Ma, J, Huang, R, Liu, J, Yao, Y, Zhang, Y & Chen, J 2016, 'Research challenges and perspectives on Wisdom Web of Things (W2T)', in Wisdom Web of Things, Springer, pp. 3-26.

Zlacky, D, Stas, J, Juhar, J & Cizmrr, A 2014, 'Text Categorization with Latent Dirichlet Allocation', *Journal of electrical and electronics engineering*, vol. 7, pp. 161-4.

6.2 Queries

All queries analysed in this study are instead in following table.

Apple -http -https –www Apple -http -https -www +exclude:retweets iPhone S6 -http -https -www iPhone S6 -http -https -www +exclude:retweets Apple -http -https -www :) Apple -http -https -www :(Apple -http -https -www +exclude:retweets :) Apple -http -https -www +exclude:retweets :(iPhone S6 -http -https -www :) iPhone S6 -http -https -www :(iPhone S6 -http -https -www +exclude:retweets :) iPhone S6 -http -https -www +exclude:retweets :(
Google -http -https Google -http -https +exclude:retweets Android -http -https Android -http -https +exclude:retweets Google -http -https :) Google -http -https :(Google -http -https +exclude:retweets :) Google -http -https +exclude:retweets :(Android -http -https :) Android -http -https :(Android -http -https +exclude:retweets :) Android -http -https +exclude:retweets :(
Samsung -http -https Samsung -http -https +exclude:retweets

```
Galaxy S6 -http -https
Galaxy S6 -http -https +exclude:retweets
Samsung -http -https :)
Samsung -http -https :(
Samsung -http -https +exclude:retweets :)
Samsung -http -https +exclude:retweets :(
Galaxy S6 -http -https :)
Galaxy S6 -http -https :(
Galaxy S6 -http -https +exclude:retweets :)
Galaxy S6 -http -https +exclude:retweets :(
```

For the sentiment and timeline analysis, following queries were used to retrieve one-day worth of Tweets:

```
iPhone -http -https -www +exclude:retweets lang:en since:2015-11-22 until:2015-11-23
iPad -http -https -www +exclude:retweets lang:en since:2015-11-22 until:2015-11-23
Apple -http -https -www +exclude:retweets lang:en since:2015-11-22 until:2015-11-23
```

The Tweets were selected for the period of November 2015 to January 2016.

6.3 Results

6.3.1 Results of the Granger causality test

The Granger causality test was performed using the R statistics programming language and the "lmtest" library. Following tables show the results as provided by the lmtest library:

6.3.1.1 Apple unclassified

Apple total unclassified	Apple deduplicated	Apple subjectivity analysis
Open quote: Model 1: Open ~ Lags(Open, 1:1) + Lags(Apple, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.0191 0.8907	Open quote: Model 1: Open ~ Lags(Open, 1:1) + Lags(Apple, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.0035 0.953	Open quote: Model 1: Open ~ Lags(Open, 1:1) + Lags(Apple, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.0061 0.9383

Close quote: Model 1: Close ~ Lags(Close, 1:1) + Lags(Apple, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.0339 0.8548	Close quote: Model 1: Close ~ Lags(Close, 1:1) + Lags(Apple, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.0177 0.8948	Close quote: Model 1: Close ~ Lags(Close, 1:1) + Lags(Apple, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.0069 0.9343

Table 6-1: Apple unclassified

6.3.1.2 Apple classified naïve Bayes

Apple total naïve Bayes	Apple deduplicated naïve Bayes	Apple subjectivity analysis naïve Bayes
Positive: Model 1: Open ~ Lags(Open, 1:1) + Lags(Positive, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.9047 0.3475	Positive: Model 1: Open ~ Lags(Open, 1:1) + Lags(Positive, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.7283 0.3988	Positive: Model 1: Open ~ Lags(Open, 1:1) + Lags(Positive, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 1.0679 0.3079
Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.3803 0.5411	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.5398 0.467	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.3992 0.5313
Negative: Model 1: Open ~ Lags(Open, 1:1) + Lags(Negative, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38	Negative: Model 1: Open ~ Lags(Open, 1:1) + Lags(Negative, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38	Negative: Model 1: Open ~ Lags(Open, 1:1) + Lags(Negative, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38

2 39 -1 0.8025 0.376	2 39 -1 0.6826 0.4139	2 39 -1 0.8025 0.376
Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.5433 0.4656	Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.7551 0.3903	Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.6855 0.4129

Table 6-2: Apple classified naïve Bayes

6.3.1.3 Apple classified Logistic Regression

Apple total Logistic Regression	Apple deduplicated Logistic Regression	Apple subjectivity analysis Logistic Regression
Positive: Model 1: Open ~ Lags(Open, 1:1) + Lags(Positive, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 2.0236 0.163	Positive: Model 1: Open ~ Lags(Open, 1:1) + Lags(Positive, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.9844 0.3274	Positive: Model 1: Open ~ Lags(Open, 1:1) + Lags(Positive, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 1.3107 0.2594
Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.4191 0.5213	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 1.375 0.2483	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.6794 0.4149
Negative: Model 1: Open ~ Lags(Open, 1:1) + Lags(Negative, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38	Negative: Model 1: Open ~ Lags(Open, 1:1) + Lags(Negative, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38	Negative: Model 1: Open ~ Lags(Open, 1:1) + Lags(Negative, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38

2 39 -1 0.6191 0.4363	2 39 -1 0.6443 0.4272	2 39 -1 0.8028 0.3759
Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.5077 0.4805	Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.5722 0.4541	Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.5759 0.4526

Table 6-3: Apple classified Logistic Regression

6.3.1.4 iPad unclassified

iPad total unclassified	iPad deduplicated	iPad subjectivity analysis
Open quote: Model 1: Open ~ Lags(Open, 1:2) + Lags(iPad, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 1.983 0.1528	Open quote: Model 1: Open ~ Lags(Open, 1:2) + Lags(iPad, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 2.1961 0.1263	Open quote: Model 1: Open ~ Lags(Open, 1:2) + Lags(iPad, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 1.2558 0.2974
Close quote: Model 1: Close ~ Lags(Close, 1:1) + Lags(iPad, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 1.9828 0.1672	Close quote: Model 1: Close ~ Lags(Close, 1:1) + Lags(iPad, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 2.433 0.1271	Close quote: Model 1: Close ~ Lags(Close, 1:1) + Lags(iPad, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 1.2541 0.2698

Table 6-4: iPad unclassified

6.3.1.5 iPad classified naïve Bayes

iPad total naïve Bayes	iPad deduplicated naïve Bayes	iPad subjectivity analysis naïve Bayes
Positive: Model 1: Open ~ Lags(Open, 1:2) + Lags(Positive, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 0.9569 0.3939	Positive: Model 1: Open ~ Lags(Open, 1:2) + Lags(Positive, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 2.1177 0.1355	Positive: Model 1: Open ~ Lags(Open, 1:2) + Lags(Positive, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 2.0221 0.1476
Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.5767 0.4523	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 2.5919 0.1157	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 2.4956 0.1225
Negative: Model 1: Open ~ Lags(Open, 1:2) + Lags(Negative, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 1.379 0.2652	Negative: Model 1: Open ~ Lags(Open, 1:2) + Lags(Negative, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 2.466 0.09955 .	Negative: Model 1: Open ~ Lags(Open, 1:2) + Lags(Negative, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 2.5918 0.08919 .
Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38	Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38	Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38

2 39 -1 1.5291 0.2238	2 39 -1 2.9943 0.09167 .	2 39 -1 3.3478 0.07515 .

Table 6-5: iPad classified naïve Bayes

6.3.1.6 iPad classified Logistic Regression

iPad total Logistic Regression	iPad deduplicated Logistic Regression	iPad subjectivity analysis Logistic Regression
Positive: Model 1: Open ~ Lags(Open, 1:2) + Lags(Positive, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 1.0166 0.3723	Positive: Model 1: Open ~ Lags(Open, 1:2) + Lags(Positive, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 2.3109 0.1141	Positive: Model 1: Open ~ Lags(Open, 1:2) + Lags(Positive, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 2.4733 0.09892 .
Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.6485 0.4257	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 2.6001 0.1151	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 3.0044 0.09115 .
Negative: Model 1: Open ~ Lags(Open, 1:2) + Lags(Negative, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 1.4699 0.2438	Negative: Model 1: Open ~ Lags(Open, 1:2) + Lags(Negative, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 2.4329 0.1025	Negative: Model 1: Open ~ Lags(Open, 1:2) + Lags(Negative, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 2.3782 0.1075

Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 1.8364 0.1834	Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 3.1578 0.08357 .	Model 1: Close ~ Lags(Close, 1:1) + Lags(Negative, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 3.167 0.08314 .

Table 6-6: iPad classified Logistic Regression

6.3.1.7 iPhone unclassified

iPhone total unclassified	iPhone deduplicated	iPhone subjectivity analysis
Open quote: Model 1: Open ~ Lags(Open, 1:1) + Lags(iPhone, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.0254 0.8743	Open quote: Model 1: Open ~ Lags(Open, 1:1) + Lags(iPhone, 1:1) Model 2: Open ~ Lags(Open, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.0253 0.8744	Open quote: Model 1: Open ~ Lags(Open, 1:2) + Lags(iPhone, 1:2) Model 2: Open ~ Lags(Open, 1:2) Res.Df Df F Pr(>F) 1 35 2 37 -2 0.7114 0.4979
Close quote: Model 1: Close ~ Lags(Close, 1:1) + Lags(iPhone, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.1495 0.7012	Close quote: Model 1: Close ~ Lags(Close, 1:1) + Lags(iPhone, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.4175 0.5221	Close quote: Model 1: Close ~ Lags(Close, 1:1) + Lags(iPhone, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 2.1619 0.1497

Table 6-7: iPhone unclassified

6.3.1.8 iPhone classified naïve Bayes

iPhone total naïve Bayes	iPhone deduplicated naïve	iPhone subjectivity analysis naïve

	Bayes	Bayes
Positive: Model 1: Open ~ Lags(Open, 1:3) + Lags(Positive, 1:3) Model 2: Open ~ Lags(Open, 1:3) Res.Df Df F Pr(>F) 1 32 2 35 -3 0.3114 0.817	Positive: Model 1: Open ~ Lags(Open, 1:4) + Lags(Positive, 1:4) Model 2: Open ~ Lags(Open, 1:4) Res.Df Df F Pr(>F) 1 29 2 33 -4 0.812 0.5278	Positive: Model 1: Open ~ Lags(Open, 1:4) + Lags(Positive, 1:4) Model 2: Open ~ Lags(Open, 1:4) Res.Df Df F Pr(>F) 1 29 2 33 -4 0.8626 0.498
Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.4017 0.53	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.1049 0.7477	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.103 0.75
Negative: Model 1: Open ~ Lags(Open, 1:4) + Lags(Negative, 1:4) Model 2: Open ~ Lags(Open, 1:4) Res.Df Df F Pr(>F) 1 29 2 33 -4 0.688 0.6061	Negative: Model 1: Open ~ Lags(Open, 1:4) + Lags(Negative, 1:4) Model 2: Open ~ Lags(Open, 1:4) Res.Df Df F Pr(>F) 1 29 2 33 -4 0.871 0.4931	Negative: Model 1: Open ~ Lags(Open, 1:4) + Lags(Negative, 1:4) Model 2: Open ~ Lags(Open, 1:4) Res.Df Df F Pr(>F) 1 29 2 33 -4 0.7886 0.5421
Model 1: Close ~ Lags(Close, 1:3) + Lags(Negative, 1:3) Model 2: Close ~ Lags(Close, 1:3) Res.Df Df F Pr(>F) 1 32 2 35 -3 0.9575 0.4247	Model 1: Close ~ Lags(Close, 1:3) + Lags(Negative, 1:3) Model 2: Close ~ Lags(Close, 1:3) Res.Df Df F Pr(>F) 1 32 2 35 -3 1.2539 0.3067	Model 1: Close ~ Lags(Close, 1:3) + Lags(Negative, 1:3) Model 2: Close ~ Lags(Close, 1:3) Res.Df Df F Pr(>F) 1 32 2 35 -3 1.3035 0.2902

Table 6-8: iPhone classified naïve Bayes

6.3.1.9 iPhone classified Logistic Regression

iPhone total Logistic Regression	iPhone deduplicated Logistic Regression	iPhone subjectivity analysis Logistic Regression
Positive: Model 1: Open ~ Lags(Open, 1:4) + Lags(Positive, 1:4) Model 2: Open ~ Lags(Open, 1:4) Res.Df Df F Pr(>F) 1 29 2 33 -4 0.5074 0.7306	Positive: Model 1: Open ~ Lags(Open, 1:4) + Lags(Positive, 1:4) Model 2: Open ~ Lags(Open, 1:4) Res.Df Df F Pr(>F) 1 29 2 33 -4 0.8829 0.4864	Positive: Model 1: Open ~ Lags(Open, 1:4) + Lags(Positive, 1:4) Model 2: Open ~ Lags(Open, 1:4) Res.Df Df F Pr(>F) 1 29 2 33 -4 0.9125 0.4698
Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.2431 0.6248	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.1027 0.7504	Model 1: Close ~ Lags(Close, 1:1) + Lags(Positive, 1:1) Model 2: Close ~ Lags(Close, 1:1) Res.Df Df F Pr(>F) 1 38 2 39 -1 0.0829 0.775
Negative: Model 1: Open ~ Lags(Open, 1:4) + Lags(Negative, 1:4) Model 2: Open ~ Lags(Open, 1:4) Res.Df Df F Pr(>F) 1 29 2 33 -4 0.6047 0.6624	Negative: Model 1: Open ~ Lags(Open, 1:4) + Lags(Negative, 1:4) Model 2: Open ~ Lags(Open, 1:4) Res.Df Df F Pr(>F) 1 29 2 33 -4 0.8075 0.5306	Negative: Model 1: Open ~ Lags(Open, 1:4) + Lags(Negative, 1:4) Model 2: Open ~ Lags(Open, 1:4) Res.Df Df F Pr(>F) 1 29 2 33 -4 0.6911 0.6041
Model 1: Close ~ Lags(Close, 1:3) + Lags(Negative, 1:3) Model 2: Close ~ Lags(Close, 1:3) Res.Df Df F Pr(>F) 1 32 2 35 -3 1.0645 0.3779	Model 1: Close ~ Lags(Close, 1:3) + Lags(Negative, 1:3) Model 2: Close ~ Lags(Close, 1:3) Res.Df Df F Pr(>F) 1 32 2 35 -3 1.3611 0.2723	Model 1: Close ~ Lags(Close, 1:3) + Lags(Negative, 1:3) Model 2: Close ~ Lags(Close, 1:3) Res.Df Df F Pr(>F) 1 32 2 35 -3 1.3087 0.2886

Table 6-9: iPhone classified Logistic Regression

6.4 Abbreviations

API	Application Programming Interface
BI	Business Intelligence
CI	Criminal intelligence
CSS	Computational social science
DJIA	Dow Jones Industrial Average
DM	Data Mining
DSS	Decision Support System
EBITDA	Earnings before interest, taxes, depreciation and amortization
FQN	Facebook query language
JSON	JavaScript Object Notation
LR	Logistic Regression
MIS	Management Information System
ML	Machine Learning
NB	Naïve Bayes
NLP	Natural Language Processing
NPV	Net Present Value
OLAP	Online Application Processing
OSN	Online Social Networks
PA	Predictive analysis
PI	Profitability Index
PV	Present value
ROI	Return on Investment
SA	Sentiment Analysis
SC	Social Capital
SM	Social Media
SMM	Social media mining
SMN	Social Media Network
SOA	Service Oriented Architecture
SVM	Support Vector Machines